Cultivar-Specific Long-Range Chromosome Assembly Permits Rapid Gene Isolation and High-Quality Comparative Analysis in Hexaploid Wheat

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Anupriya Kaur Thind

aus

Indien

Promotionskommission

Prof. Dr. Beat Keller (Vorsitz)

Asst. Prof. Dr. Simon G. Krattinger (Leitung der Dissertation)

Prof. Dr. Ueli Grossniklaus

Zürich, 2018

Table of contents

Summary

Wheat (*Triticum aestivum*) is one of the most important crops produced and consumed worldwide. Wheat production is under constant threat by a large number of biotic and abiotic stresses, of which diseases caused by pathogenic fungi are of particular importance. One of the major challenges in wheat breeding is to limit yield losses by improving biotic stress tolerance. The most sustainable strategy to improve biotic stress resistance is through the use of resistance genes. These resistance genes are used in the breeding programs either as single genes or by stacking multiple resistance genes to provide durable and broad-spectrum resistance against multiple diseases. However, only very few wheat disease resistance genes have been cloned so far and little is known about their molecular function. Therefore, it is important that we clone a large proportion of the genes that have been described, which will allow us to make informed choices for gene pyramiding to develop resistance gene cassettes. The first aim of the thesis was to develop a novel approach to rapidly isolate resistance gene in hexaploid wheat using cultivar-specific, long-range scaffolding.

To achieve this, we combined recent advances in sequencing technologies and wheat genomics, which allowed us to generate a high-quality *de novo* assembly of a single chromosome from an *Lr22a*-carrying wheat cultivar ('CH Campala *Lr22a*'). We used this novel genetic approach to isolate the wheat leaf rust resistance gene *Lr22a*.

For the cloning of *Lr22a*, we first generated a high-resolution mapping population from a cross between the susceptible Swiss spring wheat cultivar 'CH Campala' and the *Lr22a*-containing backcross line 'CH Campala *Lr22a*' and narrowed down the genetic region to 0.48 cM flanked by two SSR markers. To obtain the complete sequence across this interval, long-range *de novo* assembly was performed on flow sorted 2D chromosomes of 'CH Campala *Lr22a*' using Chicago long-range scaffolding. A 6.39 Mb scaffold was identified

that contained the two *Lr22a* flanking markers delimiting a 438 kb region. This physical interval contained nine genes and two pseudogenes. In particular, there was a cluster of two nucleotide binding, leucine-rich repeat (NLR) immune receptors and two NLR pseudogenes. A gene coding for an NLR with homology to the *Arabidopsis* RPM1 protein was validated as *Lr22a* using five EMS-derived mutants. Given its partial resistance phenotype and broad-spectrum specificity it was very surprising that *Lr22a* codes for an NLR because most known NLRs provide complete and race-specific resistance. Therefore, *Lr22a* might reveal a novel molecular mechanism of NLR function.

This novel approach of gene cloning overcomes the limitations of traditional map-based cloning, which requires repeated rounds of chromosome walking and Bacterial Artificial Chromosome (BAC) sequencing to get sequence information from a donor line containing the gene of interest. The major limitation of BAC clones is their short insert size of 100-200 kb.

The availability of high quality reference sequences of a particular species has made significant contributions to the understanding of genome structure. However, gene order and content, as well as gene sequences can differ dramatically between accessions of the same species. This brings us to the second objective of the thesis which was the comparative analysis of the high-quality sequences of two wheat cultivars, Chinese Spring and 'CH Campala *Lr22a*', to determine genomic differences. We conducted a megabase-scale chromosome sequence comparison of the 2D chromosome and identified four large InDels, two of which showed copy number variation for NLR immune receptors. We predicted the precise breakpoints and the underlying mechanism for two of the four InDels. Apart from InDels, we also found four diverse haploblocks of ~4 Mb, ~8 Mb, ~9 Mb and ~48 Mb with a 35-fold increased SNP density compared to the rest of the chromosome. Gene comparison

between the two cultivars revealed that 99% of the genes were conserved with only 0.43 to 0.73% of unique genes (genes which are only present in one cultivar).

Our study highlights the significance of using high-quality sequences to determine large structural variations that are more than 100 kb in size. Most of the previous studies in wheat were based on short-read sequences, which were highly fragmented and incomplete and could therefore only reveal small structural variations. Our comparative analysis forms the basis for future pan-genome studies of multiple, high-quality wheat assemblies.

Zusammenfassung

Weizen (*Triticum aestivum*) ist eine der am meisten produzierten und konsumierten Nutzpflanzen weltweit. Weizen ist einer grossen Zahl an biotischen und abiotischen Formen von Stress ausgesetzt, unter denen pathogene Pilze von spezieller Relevanz sind. Eine der Hauptherausforderungen in der Weizenzüchtung ist die Reduktion von Ertragseinbussen durch Verbesserung der Toleranz gegenüber biotischem Stress. Im Laufe der Evolution haben Pflanzen verschiedene Mechanismen entwickelt, um biotischem Stress entgegenzuwirken. Die nachhaltigste Strategie zur Verbesserung der biotischen Stressresistenz ist die Nutzung von Resistenzgenen. Diese Resistenzgene werden in Züchtungsprogrammen entweder einzeln oder durch Pyramidisierung mehrerer Resistenzgene verwendet, um dauerhafte und breitgefächerte Resistenz gegen verschiedene Krankheiten zu gewährleisten. Jedoch sind bisher nur wenige Krankheitsresistenzgene von Weizen kloniert worden und nur wenig ist bekannt über deren molekulare Funktion. Deshalb ist es wichtig, dass ein grosser Anteil der beschriebenen Gene kloniert wird, was es uns ermöglicht, eine sinnvolle Auswahl für die Pyramidisierung von Genen zur Entwicklung von Resistenzgenkassetten zu treffen. Das erste Ziel meiner Doktorarbeit war es, eine effiziente Methode zu entwickeln, Resistenzgene schnell aus hexaploidem Weizen mittels Kultivar-spezifischem „long-range Scaffolding “ zu isolieren.

Um dies zu erreichen, kombinierten wir kürzlich erzielte Fortschritte in Sequenziertechnologien und der Weizengenetik, welche es uns erlaubten, ein *de novo* Chromosom „Assembly“von hoher Qualität des *Lr22a*-tragendem Weizenkultivars (‘CH Campala *Lr22a*’) zu generieren. Wir nutzten diesen neuen genetischen Ansatz zur Isolierung des Braunrostresistenzgens *Lr22a* von Weizen. Für die Klonierung von *Lr22a* kreierten wir eine hochauflösende Kartierungspopulation durch die Kreuzung des anfälligen Schweizer

Sommerweizens `CH Campala` und die *Lr22a*-tragende Rückkreuzungslinie „*Lr22a* Campala“ und engten die genetische Region durch zwei SSR Marker auf 0.48 cM ein. Um die komplette Sequenz dieses Intervalls zu erhalten, wurde ein *de novo* „long-range Assembly“ ausgeführt mit „flow-sorted“ 2D Chromosomen von ‘CH Campala *Lr22a*’ mittels Chicago „long-range Scaffolding“. Es wurde ein 6.39 Mb „Scaffold“ identifiziert, welches die zwei *Lr22a* flankierenden Marker einschliesst. Die flankierenden Marker lokalisieren das Gen in einen Bereich von 438 kb. Dieses physikalische Intervall enthält neun Gene und zwei Pseudogene. Insbesondere gibt es einen Bereich bestehend aus zwei „nucleotide binding, leucine-rich repeat (NLR)“ Immunrezeptoren und zwei NLR Pseudogenen. Ein NLR mit Homologie zu dem *Arabidopsis* RPM1 Protein wurde mittels fünf EMS-Mutanten validiert. Es ist überraschend, dass *Lr22a* eine partielle Resistenz mit breitem Wirkungsspektrum auslöst, da NLRs üblicherweise für komplette und Rassen-spezifische Resistenz bekannt sind. Darum könnte *Lr22a* zur Entdeckung eines neuen molekularen Mechanismus der NLR Funktion beitragen.

Der in dieser Dissertation beschriebene Ansatz der Genklonierung überwindet die Limitierungen der traditionellen kartengestützten Kartierung, welcher mehrere Wiederholungen des Kartierens entlang des Chromosoms und die Sequenzierung von Bacterial Artificial Chromosomes (BACs) erfordert, um die Sequenzinformation der Donorlinie mit dem Gen von Interesse zu erhalten. Die Haupteinschränkung der BAC Klone ist die kleine Grösse (100-200 kb) des inserierten DNA Stücks.

Die Verfügbarkeit von Referenzsequenzen in hoher Qualität einer bestimmten Spezies ist von grosser Wichtigkeit zur Entschlüsselung von Genomstrukturen. Jedoch können sich Genreihenfolge, -dichte sowie -sequenzen zwischen Akzessionen der gleichen Art dramatisch unterscheiden. Dies resultiert in dem zweiten Ziel der Dissertation, eine vergleichende Analyse qualitativ hochwertiger Sequenzen zweier Weizenkultivare, Chinese Spring und ‘CH

Campala *Lr22a*', zur Bestimmung der genomischen Unterschiede durchzuführen. Wir führten im Megabasenbereich einen Sequenzvergleich der 2D Chromosomen durch und identifizierten vier grosse InDels, von denen zwei sich als eine Variation in der Kopienanzahl der NLR Immunrezeptoren herausstellten. Wir konnten die zwei exakten Bruchpunkte und zugrundeliegenden Mechanismen für zwei der vier InDels bestimmen. Abgesehen von den InDels fanden wir auch vier unterschiedliche Haploblocks von ~4 Mb, ~8 Mb, ~9 Mb and ~48 Mb mit einer 35-fach erhöhten SNP-Dichte verglichen mit dem Rest des Chromosoms. Der Genvergleich zwischen den zwei Kultivaren deckte auf, dass 99% der Gene konserviert waren mit nur 0.43 bis 0.73% einmaligen Genen (Gene, welche nur in einem Kultivar anwesend sind).

Unsere Studie unterstreicht die Wichtigkeit von Sequenzen hoher Qualität, um grosse strukturelle Unterschiede von mehr als 100 kb zu bestimmen. Die meisten der bisherigen Studien in Weizen basierten auf „short-read" Sequenzen, welche stark fragmentiert und nicht vollständig waren und deswegen nur kleine strukturelle Variationen aufdecken konnten. Unsere vergleichende Analyse bietet eine Basis für zukünftige pan-genomische Studien mehrerer Weizengenome von hoher Qualität.

Chapter 1

General Introduction

1.1 Wheat genome evolution

Wheat was one of the first crops to be domesticated 10,000 years ago and it is adapted to the temperate regions of the world. The grains of wheat provide a rich source of proteins, carbohydrates and minerals. Wheat serves as a staple food for 40% of the world's population and accounts for 20% of the caloric intake (Saintenac et al., 2018). The rise of modern agriculture and wheat domestication played a major role in the shaping of human history. Initial farming practices made use of the wild diploid wheat species such as *Triticum* species but as agriculture advanced, these diploid wild species were substituted with domesticated diploid and polyploid wheat species (Salamini et al., 2002). There are two major types of cultivated polyploid wheat; the tetraploid pasta wheat (*Triticum turgidum* ssp. *durum*; AABB genomes) and the hexaploid bread wheat (*Triticum aestivum*, AABBDD genomes), latter dominating the global wheat production. *T. aestivum* has a genome size of 15.8 Gb and a repeat content of more than 85% (International Wheat Genome Sequencing Consortium, 2018; Wicker et al., 2018).

The genus *Triticum* was divided into three taxonomic groups: einkorn, emmer and bread wheat. The three groups differ in their chromosome number: einkorn wheats are diploid with 2n=2x=14, emmer wheat is tetraploid with 2n=4x=28 and bread wheat is hexaploid with 2n=2x=42. The hexaploid wheat arose from two polyploidization events. The first one occurred 0.58 to 0.82 million years ago where 7 chromosomes pairs of the diploid wheat having genome A (*T. urartu*) hybridized with a diploid wheat with genome B to constitute the 14 chromosomes of tetraploid wheat called wild emmer wheat (*T. turgidum* ssp *dicoccoides*, AABB). This wild emmer was put under cultivation where domestication and selection led to the formation of the cultivated emmer (*T. turgidum* ssp *dicoccum*, 2n=2x=28, AABB) from which free-threshing tetraploid, *T. durum*, evolved (Dvorak et al., 2012).

It was known since the 1920s that *T. turgidum* was the source of the AABB genome of hexaploid wheat, but the origin of the D genome donor remained unknown until 1940. It was then identified that *Ae. tauschii* is the donor of the D genome, based on artificial crosses between *T. dicoccoides* and *Ae. squarrosa* (also known as *Ae. tauschii*) that led to amphidiploid wheat (Mcfadden & Sears, 1944). Thus, in a second polyploidization event 14 chromosome pairs of cultivated emmer were combined with the 7 pairs of genome D (*Ae. tauschii*, DD), leading to the formation of the 21 pairs of the modern hexaploid bread wheat (*T. aestivum*, 2n=6x=42, AABBDD) (Huang et al., 2002).

The progenitors of the A and D genome of hexaploid wheat were identified based on the high degree of homology with the diploid genome of the wild einkorn *T. urartu* (2n=14, AA) and the wild goatgrass *Ae. tauschii* or *Ae. squarrosa* (2n=14, DD), respectively. In contrast, the progenitor of the B genome is highly debated. It is speculated that the progenitor donor species of the B genome might have undergone massive differentiation later, making it difficult to identify the origin of the hexaploid wheat B genome (Feldman & Levy, 2005). Another hypothesis for the ambiguous nature of B genome is that the donor of the B genome exists but has not been found yet or is extinct (Feldman & Levy, 2005). However, various morphological, cytological, biochemical, geographical and molecular studies have revealed that the species of section *Sitopsis* of *Aegilops* whose closest modern relative is *Ae. speltoides* (2n=14 genome SS) could have been the donor of the B genome (Feldman & Levy, 2005).

1.2 Human population and agriculture

The human population is expected to increase to 10 billion by 2050, which will result in significantly increased demand for food (http://www.un.org/en/development/desa/news/population/2015-report.html). In order to meet the demands of the increasing population, it is important to increase the production of

cereal crops by 110% in the next 50 years (Tilman et al., 2011). Cereal crops, such as wheat, rice, maize, barley and rye play an important role in the global food production, of which, wheat serves as a staple food for 40% of the world population (Peng et al., 2011). The global wheat production for the year 2016/17 amounted to 761.34 million tons and ranked second after maize (1040.37 million tonnes) (FAO, 2017). Hexaploid wheat accounts for the majority (95%) of the global wheat production and the remaining 5% is constituted by the tetraploid durum wheat (FAO, 2017). To account for the increasing demand of wheat, it is important to increase the yield potential and the actual yield to ensure sufficient food. Yield potential refers to the yield of a cultivar achieved when grown under optimal agronomical practices and free from biotic and abiotic stress. Actual yields in the field are usually lower than the potential yield that could be achieved under optimal conditions, which is due to yield losses caused by various biotic and abiotic stresses. Biotic stresses are caused by living organisms such as fungi, viruses, bacteria, oomycetes, insects and weeds, whereas abiotic stress is caused by non-living agents such as heat, cold and drought. The major challenge in wheat breeding is to limit the yield loss by improving biotic and abiotic stress resistance. Biotic stresses alone cause on average 10-15% yield loss in cereals (Chakraborty & Newton, 2011; Fisher et al., 2012). In wheat, 10% of the yield loss was attributed to biotic stresses in Central Africa, Southeast Asia, America and Northwest Europe for the year 2001-2003 (Oerke, 2006).

1.3 Diseases of wheat

Wheat is under constant attack from various fungal, viral and bacterial pathogens commonly found in most agricultural systems. The key diseases caused by these pathogens are listed below.

The most important among all the diseases are the ones caused by fungal pathogens. Based on their lifestyles, fungal pathogens can be classified as biotrophic, necrotrophic and

hemi-biotrophic. Biotrophic pathogens proliferate on living plant tissue whereas necrotrophic pathogens feed on dead plant material. Hemi-biotrophic pathogens combine a biotrophic growth phase usually during early stages of infection followed by a necrotrophic phase. Examples of biotrophic fungal pathogens include *Puccinia triticina, P. striiformis and P. graminis* which cause the three rust diseases of wheat, namely leaf rust, stripe rust and stem rust respectively. *Blumeria graminis* is another important biotrophic pathogen and causes powdery mildew (Saari & Prescott, 1985) which is the most common foliar disease of wheat and is characterised by powdery white to grey fungal growth on leaves, stems and head. Examples of hemi-biotrophic fungal pathogens are *Fusarium graminearum* and *Zymoseptoria tritici* which cause Fusarium head blight (FHB) and Septoria tritici blotch (STB), respectively. *F. graminearum* starts to infect the open florets as a biotrophic pathogen where extracellular hyphae grow in the living host cell without visible disease symptoms and later switches to the nectrophic phase when the fungal hyphae enter the wheat cells which is followed by host cell death. FHB is a hazardous floral disease of wheat globally which caused yield loss of approximately US$ 3 billion between 1990a and 2008 in the US alone (Figueroa et al., 2017). FHB affects grain yield, quality and also results in the accumulation of mycotoxins in grain which poses a major food safety risk. *Zymoseptoria tritici* is a latent necrotroph with a latent asymptomatic phase during which it grows and protects itself from plant defences prior to switching to a strong nectrophic growth (Saintenac et al., 2018). STB causes substantial grain yield and quality losses under favourable environment and is a devastating wheat disease in Europe, which leads to 5-10% annual wheat loss. STB is primarily managed by fungicides, although there are 21 major STB resistance genes that have been genetically defined and to date only one (*Stb6*) has been cloned (Saintenac et al., 2018).

An example of a necrotrophic pathogen is *Bipolaris sorokiniana* (Bs), the causal agent of leaf blight. Bs is a devastating pathogen that causes both foliar and root diseases and is a major biotic constraint in wheat growing areas of Eastern Gangetic plains in India, Bangladesh and

where the losses can reach up to 50% under favourable conditions (Figueroa et al. 2017). The molecular basis of the infection is currently unknown as only a handful of genes have been characterized (Figueroa et al., 2017) from the pathogen and no avirulence gene (*Avr*) has been identified so far.

1.3.1 Wheat rust diseases

Wheat rusts caused by fungal pathogens belonging to the genus *Puccinia* are the most serious biotic constraints for wheat production worldwide. Wheat rust fungi are obligate pathogens that require an alternate host (*Berberis vulgaris* (barberry shrub) for stem and stripe rust (Jin, 2011) and *Thalictrum speciosissimum* for leaf rust) (Bolton et al., 2008) for completing their sexual life cycle. In general, the alternate hosts are required for the rust pathogens to diversify rust populations through sexual reproduction and overcome resistance in wheat and barley. For example, a heavily infected barberry shrub can give rise to 70 billion genetically diverse spores (Schwessinger, 2017). To reduce stem rust epidemics, barberry plants were eradicated in United States from1920 through the 1970 and this eradication program significantly reduced stem rust epidemics in major wheat producing areas and reduced the number of races (Wang et al., 2015). Eradication of the alternate host species therefore plays an important role in protecting the cereals against these rust diseases (Kolmer et al., 2007).

Rusts are termed biotrophs because they extract nutrients from the living host cells. Wheat stem rust is caused by *P. graminis* f. sp. *tritici* and is widely distributed around the world, although less common than stripe rust and leaf rust (Singh et al., 2015). It affects leaf sheath, stem, glumes and awns of susceptible plants leading to global yield loss of average 6.2 million metric tons per year (Pardey et al., 2013). However, stem rust has gained significant importance in the recent years due to the emergence of new *Pgt* isolates which affects wheat cultivars around the globe. The emergence of Ug99 in Uganda and its expansion across East Africa, the Middle

East and the appearance of its variants is posing a serious threat to wheat because 90% of the wheat varieties in the world are susceptible to Ug99 (Singh et al., 2015). Also, recent re-emergence of stem rust in Europe, mainly in Germany, Sicily, Sweden and UK is becoming an increasing threat for wheat production which calls for re-initiation of resistance breeding and eradication of the alternate host near wheat growing areas (Lewis et al., 2018).

Stripe rust is caused by *P. striiformis* f. sp. *tritici* (*Pst*). It is currently the most economically important rust disease that can lead to yield loss of 100% if a susceptible cultivar is grown (Chen, 2005). Approximately, 88% of the wheat cultivars have been reported to be susceptible to *Pst* leading to a global loss of US$ 1 billion annually (Beddow et al., 2015). The threat of this fungus to agriculture is mainly due to huge genetic diversity because of sexual recombination in the Himalayan region and its dispersal across continents (Schwessinger, 2017). Genetic control of stripe rust can be achieved by the use and combination of more than 50 resistance genes identified over the last 100 years (https://wheat.pw.usda.gov/GG2/Triticum/wgc/2013/).

Leaf rust is the most common and widely distributed wheat rust disease and is caused by the *P. triticina* (*Pt*). Leaf rust infection can lead to reduction in kernel weight and the number of grains per head (Figueroa et al., 2017). The yield losses caused by leaf rust display a geographical and temporal variation (Huerta-Espino et al., 2011). Leaf rust is a problematic disease because the pathogen displays high diversity as there is constant emergence of new races that are adapted to wide range of climates (Huerta-Espino et al., 2011; McCallum et al., 2016). Leaf rust is discussed in detail in the following section.

1.3.2 Leaf Rust – life cycle and economic importance

P. triticina, the causal agent of leaf rust completes its life cycle on two unrelated hosts and hence the name heteroecious rust. Leaf rust is macrocyclic and has five different stages of

which teliospores, basidiospores and urediniospores are produced on the cereal host and the other two stages, pycniospores and aeciospores, are on the alternate host (Kolmer, 2013). The urediniospores are dikaryotic and can re-infect the plant host when water is present on the leaf surface and temperatures are between 10 and 25°C. As host plants mature, teliospores are produced in the uredinia and the dikaryotic nuclei undergo karyogamy to produce a diploid nucleus.

Under suitable conditions, one or both cells in the teliospores produce a hyphal protrusion called promycelium. The diploid nucleus undergoes meiosis and the four haploid nuclei migrate into the promycelium where they are divided into four different cells (each with one haploid nucleus) by septa. A spike-like structure is formed on the apical wall of these four cells and the haploid nucleus migrates into the newly formed basidiospore. The nucleus in the basidiospore undergoes mitosis and forms a mature single-cell basidiospore. The basidiospores are carried by the air from the host plant and then land on the upper leaf surface and infect the alternate host (*Thalictrum speciosissimum*) (Fig. 1.1) leading to the development of pycnial structures. Inside these pycinial structures, haploid pycniospores and flexous hyphae are produced which function as male and female gametes, respectively, and these are carried by insects to other pycinial infections where they combine with the opposite mating type to form an aecium and the dikaryotic nuclear condition is restored. On maturation, aecia release aeciospores and these aeciospores infect the cereal host, thus completing the life-cycle. *Pt* can cycle indefinitely on plant host as uredinial infections (Fig. 1.1).

The rust pathogens form specialised infection structures to invade host. These structures are required for spore attachment, host recognition, penetration, proliferation and nutrition (Mendgen & Hahn, 2002). The most complex infection structure is the haustorium, which serves as a feeding structure to extract nutrients from the host and also to supress the defence responses

triggered by the plant (Fig. 1.2). The infection starts with the landing of the spores on the host plant followed by the germination of these spores. This germination process leads to the development of the primary germ tube, whose growth is directed towards the leaf stomata by thigmotrophic growth and results in the formation of the appressorium over the stomatal aperture. This promotes the growth of the infection peg into the substomatal cavity and formation of the infection hyphae (Fig. 1.2). On contact with the mesophyll cells, these infection hyphae form haustoria mother cells, which further leads to the development of haustoria. The haustoria later invaginate the mesophyll plasma membrane by penetrating through the cell wall (Webb & Fellers, 2006). The orange-coloured infection pustules called urediniospores appear 7-10 days post infection (Fig. 1.3) and new infection cycle starts with the water or wind dispersed urediniospores.

The fungal disease leaf rust is the most common and most widespread rust disease, occurring in all wheat growing areas (Bolton et al., 2008; Kolmer, 2013). Early infection of the flag leaf (60-70% of leaf covered with pustules) during spike emergence can cause more than 30% of yield loss (Huerta-Espino et al., 2011). However, if the same level of infection occurs at the soft dough stage, the yield loss can be comparatively lower (7%) (Huerta-Espino et al., 2011). In Canada, a yield loss of 10% was reported between 2000-2009 and in South America during 1999-2003, a yield loss of more than 50% was observed under favourable leaf rust conditions (Huerta-Espino et al., 2011). In India, Pakistan, Bangladesh and Nepal, leaf rust caused yield loss of 10-25% whereas in US, a yield loss caused by *Pt* was estimated to be over US$ 350 million for the year 2000-2004 (Huerta-Espino et al., 2011).

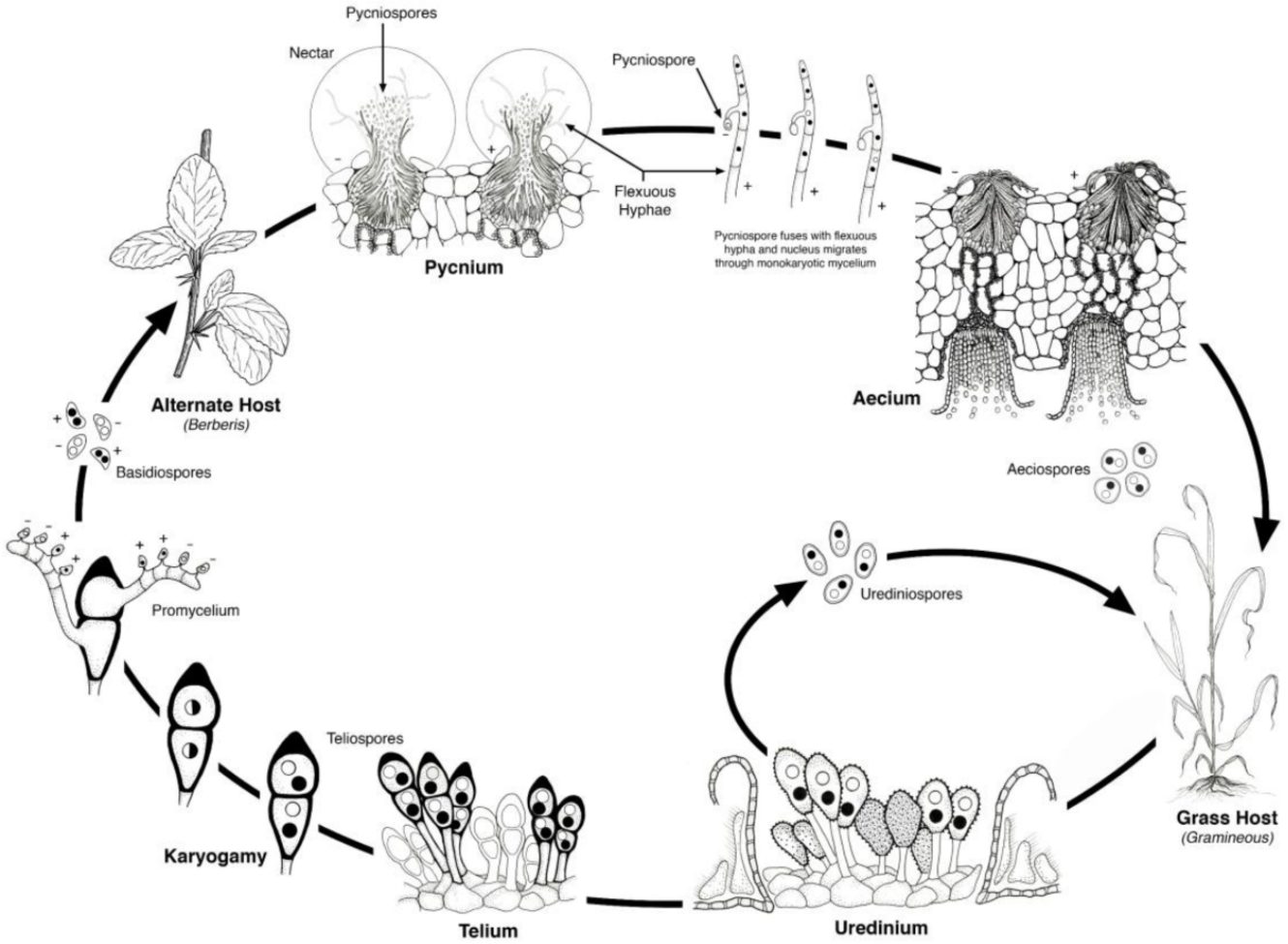


Fig. 1.1 Life cycle of the basidiomycete fungus, *Puccinia triticina* (source: Leonard and Szabo, 2005).

Fig. 1.2 Wheat leaf rust visible as medium to large sized, orange coloured uredinia.

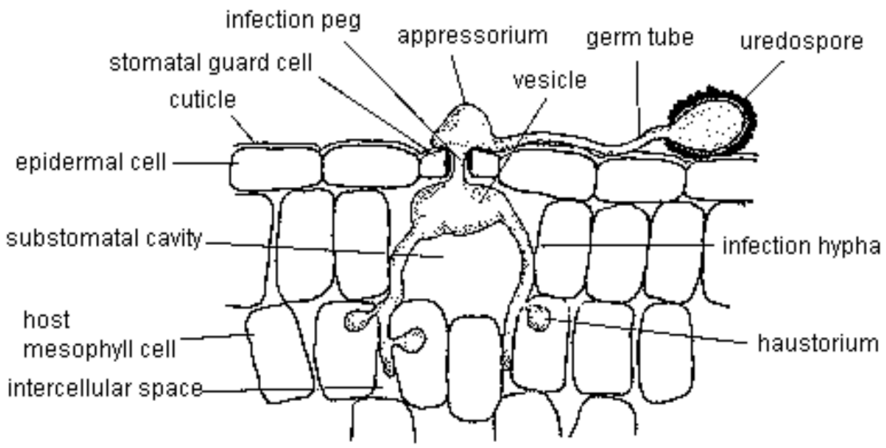


Fig. 1.3 Diagrammatic representation of the infection structures produced by rust fungi during host leaf invasion (source: http://web.unbc.ca/ctl/webcourses/fsty307-2/rust.html)

1.4 Disease resistance in Crop Plants

Disease resistance in plants can be classified into various categories based on phenotypic or genetic characteristics (Table 1.1).

Table 1.1 Phenotypic and genetic characterization of resistance genes

Categories	Types	Definition
Durability	Durable resistance genes	Resistance that has remained effective in a cultivar during its widespread cultivation over a long period of time and in an environment favourable to a disease (Johnson, 1984)
	Non-durable resistance gene	Resistance that has been overcome by certain pathogen races.
Specificity	Race-specific	Operates against few races of a given pathogens.
	Race-non-specific	Operates against all races of a given pathogen.
Stage of resistance	Seedling resistance (all stage resistance)	Resistance provided at all the stages from seedling to the adult plant
	Adult plant resistance (APR)	Resistance at the adult plant stage only
Nature of genetic control	Monogenic	Resistance is controlled by single gene
	Polygenic	Resistance to several, different races is controlled by multiple genes.
Level of Infection	Partial resistance	Disease progresses at retarded rate and as a result it shows low or intermediate level of resistance (Vale et al., 2001)
	Complete resistance	Resistance that does not allow growth of the pathogen. There are no signs of disease or of the presence of the pathogen

The following paragraphs will provide an overview of the defense mechanisms in plants and approaches used for rust resistance breeding in wheat.

1.4.1 Immunity in Plants

Plants lack an adaptive immune system comparable to that found in humans and they solely rely on innate immune mechanisms (Jones & Dangl, 2006). A great proportion of the immunity in plants is encoded by genes producing immune receptors that can be broadly classified into two categories based on their subcellular localization: 1) plasma membrane-localized receptors with an extracellular ligand-binding domain; and (2) intracellular immune receptors (Cook et al., 2015; Dodds & Rathjen, 2010; Jones & Dangl, 2006; Thomma et al., 2011). These receptors perceive pathogen or host-derived signatures that are produced during pathogen penetration (Krattinger & Keller, 2016). Most intracellular immune receptors identified so far belong to the nucleotide binding-leucine-rich repeat receptor (NLR) family. A typical plant genome contains several hundred NLR genes (Sarris et al., 2016). NLRs directly or indirectly perceive virulence effectors proteins that are released by the pathogen during the infection process. Virulence effectors target and modify plant proteins that are involved in basal defense mechanisms. Activation of NLR proteins by effector proteins induces a defense response consisting of series of cellular and biochemical processes and transcriptional reprogramming, which often leads to programmed cell death called hypersensitive response (HR) (Belkhadir et al., 2004; Dangl & Jones, 2001; Nimchuk et al., 2003).

NLR proteins can bind directly to the effector proteins (Fig. 1.3). For example, genetic studies of flax and flax rust (*Melampsora lini*) interactions showed a direct interaction between the identified NLR gene in host and the virulence effector in rust pathogen. *R* genes cloned from four of the five loci in flax encoded for TIR-NLR class and effector genes cloned from four loci of rust encoded for small secreted proteins with no similarity on the amino acid level between

the cloned effectors. It was observed that the flax NLR L567 directly perceives the flax rust effector AvrL567 leading to activation of the resistance defense response (Ellis et al., 2007).

However, for many known NLR proteins, a direct physical interaction with the effector protein could not be experimentally demonstrated. This led to the hypothesis that the interaction can also be indirect as shown in Fig. 1.3 (van der Hoorn & Kamoun, 2008). In this model, a NLR protein monitors the status of an effector target, which is often a component of the basal immunity. Changes to this 'guarded' molecules are perceived by the NLR, which triggers a defense response (Dangl & Jones, 2001; van der Biezen & Jones, 1998). The indirect recognition model (or "guard model") explains how a relatively small number of NLRs guarding a finite number of host targets can confer resistance to multiple pathogens with an almost infinite capacity for effector variation (Dangl & Jones, 2001). One of the classical examples of a guarded immune protein is the *Arabidopsis* RPM1 interacting protein 4 (RIN4), which is a master switch of basal defense. RIN4 is targeted by multiple *Pseudomonas syringae* effectors (AvrRpm1, AvrRpt2 and AvrB) that alter its activity to promote infection (Bisgrove et al., 1994; Grant et al., 1995). Two NLR proteins RPM1 and RPS2 guard RIN4 (Kim et al., 2009). The effectors AvrB and AvrRPM1 from *P. syringae* lead to the phosphorylation of the RIN4. The perturbation of RIN4 by effectors induces the activity of the R protein leading to resistance but there are no reports if and how these modifications benefit pathogen virulence (van der Hoorn & Kamoun, 2008). Another effector, AvrRpt2 which is a putative cysteine protease (Axtell et al., 2003) leads to posttranscriptional disappearance or cleavage of RIN4 (Mackey et al., 2002). This disappearance of RIN4 activates RPS2, generating a resistance response (Belkhadir et al., 2004; Kim et al., 2009).

Some NLRs carry integrated domains (NLR-IDs) that act as effector trap. For example, *RGA5* and *Pik-1* are two rice NLRs which have an additional heavy metal-associated domain

(HMA). The HMA domain recognizes several rice blast effectors and triggers a defense response (Cesari et al., 2013; Maqbool et al., 2015). It is assumed that HMA domain-containing proteins are components of the basal immune response in rice and that they are consequently targeted by pathogen effectors (Krattinger & Keller, 2016). Another study based on the whole genome comparison of 40 publically available plant genomes by Sarris et al (2016) identified that 'integrated domain' NLRs are frequently present. They found a total of 265 distinct integrated domains in 750 NLR proteins. Some of them such as WRKY and protein kinase domain have a known function in basal immunity, however, for some of these integrated domains no link to plant immunity has been found so far.

The close genetic interaction between an NLR and its corresponding effector imposes a reciprocal, antagonistic selection pressure on the effector to escape NLR detection and the NLR to maintain effector recognition (Aguileta et al., 2009). Therefore, diversifying selection, acting on variation generated by recombination and mutation, has often resulted in great diversity between NLR loci making it difficult to discern orthogonal relationships between haplotypes from the same species (Jacob et al., 2013; Parniske et al., 1997). Two evolutionary models have been proposed to explain NLR-resistance gene diversification. In the "arms race" model, a new virulent allele of the pathogen effector gene evolves to avoid recognition by the host resistance gene (Kanzaki et al., 2012; Woolhouse et al., 2002). The frequency of this allele increases in the pathogen population because of fitness advantage and eventually replaces the old allele. In turn, a new host *R* gene allele evolves to recognize the pathogen and prevent infection. The new host allele increases in frequency until fixed in the population. Thus, the cycle of birth and death continues.

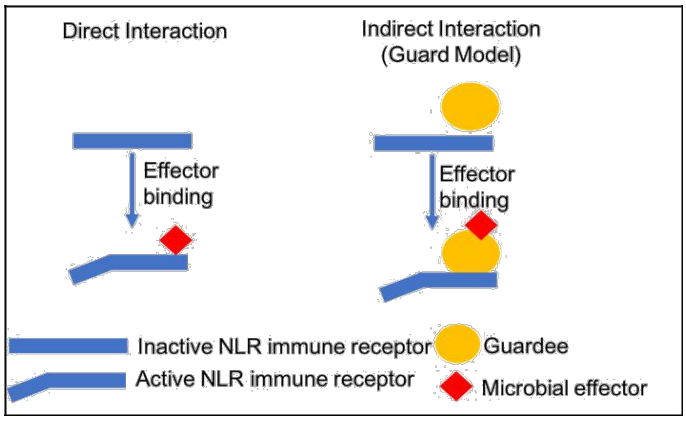


Fig. 1.3 NLR-mediated direct and indirect perception of virulence effectors. During indirect effector perception, a NLR monitors the status of an effector target and changes to the 'guarded' protein activates NLR-mediated resistance (Image modified from Krattinger and Keller, 2016).

In the second model, "trench warfare" (Stahl et al., 1999), the allele frequencies of the corresponding genes in host and pathogen fluctuate because of negative frequency-dependent selection (Woolhouse et al., 2002). This results in the maintenance of the same set of alleles over greater time periods with more within-species polymorphism, while resistance genes following the "arms race" trajectory tend to display more inter-species polymorphism (Woolhouse et al., 2002).

1.4.2 Rust resistance genes in wheat – Nucleotide binding-leucine-rich repeat receptors (NLRs)

NLRs provide qualitative disease resistance. Qualititative resistance shows a discontinuous range of variation in resistance in the host genotype and susceptible and resistant genotypes can be easily discerned. In wheat and its wild relatives, more than 150 rust resistance genes have been genetically defined (McIntosh et al., 1995). To date, 14 rust resistance genes have been cloned of which 11 namely, *Sr13*, *Sr22*, *Sr45*, *Sr50*, *Sr33*, *Sr35*, *Yr10*, *Lr1*, *Lr21*, *Lr10,* and *Lr22a* encode for NLRs (Ellis et al., 2014; Mago et al., 2015; Periyannan, 2013; Saintenac et al., 2013; Steuernagel et al., 2016; Thind et al., 2017). This class of genes is extensively used in breeding programs and some NLR genes have been used in many wheat cultivars that are grown

on large acreages worldwide. The extensive use of certain NLR genes fostered the rapid adaptation of the pathogen and hence, breakdown of disease resistance. (Krattinger & Keller, 2016). An example of the breakdown of a resistance gene is *Sr31,* which was translocated from rye (*Secale cereale* L.) to wheat on a large chromosomal translocation of rye chromosome 1RS. *Sr31* was effective against all *Pgt* races for 30 years until the emergence of African stem rust race Ug99 (Pretorius, 2000) which led to susceptibility in 90% of the wheat cultivars worldwide amounting to 20% of the yield loss in Asia, Middle East, Central and North Africa (Singh et al., 2011). Another example is *Sr24,* which was introgressed from *Agropyron* sp. and is a wheat stem rust resistance gene which was overcome by the virulent pathogen races globally except in Australia (McIntosh et al., 1995).

For leaf rust, four resistance genes have been cloned and all encode for NLR. However, they do not share any similarity either at the DNA or at the protein level. *Lr21* is located on chromosome arm 1DS of bread wheat and was introgressed from *Ae. tauschii* using a synthetic wheat. Sequence analysis revealed that *Lr21* is a chimera of two non-functional *lr21* haplotypes (H1 and H2) (Huang et al., 2003). Bread wheat only contains the inactive *lr21* alleles, however, an active allele can be experimentally reconstituted by recombination of inactive H1 and H2 (Krattinger & Keller, 2016). In contrast to *Lr21*, *Lr1* was first described in hexaploid wheat cultivar Malakoff (Dyck & Samborski, 1968). *Lr1* is located on chromosome 5D of wheat. Sequence analysis of *Lr1* revealed a specific polymorphism of 605 bp, encoding LRRs 9-15 which distinguished *Lr1* from its susceptible allele (Cloutier et al., 2007). *Lr10* is a single copy gene located on chromosome 1AS (Feuillet et al., 2003). *Lr10*-mediated leaf rust resistance is highly unusual because of its dependence on two CC-NBS-LRR (encoded by *Lr10* and *RGA2*) in tetraploid and hexaploid wheat and also because of its pattern of diversifying selection in the LRR domain (Loutre et al., 2009). Typically, the LRR domain is under diversifying selection, as in cases of direct interaction with the pathogen effector, the LRR domain confers specificity to

the interaction (Dodds et al., 2001). The N-terminal domain is usually highly conserved in an allelic series of *R* genes, for example, *Mla* and *Pm3* powdery mildew resistance genes in barley and wheat, respectively, both have conserved N-terminal domains (Bieri et al., 2004; Yahiaoui et al., 2006). *Lr10*, therefore indicates a complex mechanism of pathogen recognition and signal transduction. However, there is now evidence that the paired NLRs form heterogenous protein complexes even during the initial signal activation (Williams et al., 2014) where the first member of the protein is involved in binding the pathogen effector, guardee or an integrated domain and the second member is involved in sensing the changes in the first NLR and initiating signalling (Baggs et al., 2017).

1.4.3 Rust resistance genes in wheat – Non-Nucleotide binding-leucine-rich repeat receptors (Non-NLRs)

There is another class of resistance genes that provides 'quantitative' disease resistance. Quantitative resistance (QR) is partial and is usually provided by the joint effect of 3-5 additively acting genes. Although there is no a priori connection between the completeness of resistance genes' action and their durability, it has been found that QR genes are often more durable than NLR genes (Ellis et al., 2014). The durability aspect of QR genes is difficult to study in model plants. Consequently, most of our knowledge on QR genes has been derived from crop plants. Though QR is usually provided by the joint effect of several genes in one cultivar, single QR genes can be mendelized as individual quantitative trait loci (QTL) and their phenotypic effects assessed as single genes. For example, *pi21*, a recessive rice blast resistance gene that encodes for a proline-rich protein was identified as one of several QR gene in a QTL study (Fukuoka & Okuno, 2001). Through repeated backcross, *pi21* was transferred to a near-isogenic background that allowed assessment of *pi21* as single gene (Fukuoka et al., 2009).

Of the 14 cloned rust resistance genes in wheat, three are non-NLRs, namely *Lr34*, *Lr67* and *Yr36* (Fu et al., 2009; S. G. Krattinger et al., 2009; Moore et al., 2015). *Lr34* and *Lr67* both encode for membrane-localised transporter proteins. *Lr34* is one of the most widely used adult plant resistance (APR) genes in wheat breeding. It shows a partial resistance phenotype and has been used in commercial wheat cultivars for 100 years (Johnson, 1988). *Lr34* was cloned and it encodes for an ABC-transporter family protein (Krattinger et al., 2009). *Lr34* was found to be completely linked to *Yr18* (partial yellow rust resistance), *Pm38* (partial powdery mildew resistance) and *Sr57* (partial stem rust resistance) and shows a leaf tip necrosis phenotype. Several independent mutants that had a point mutation within the *Lr34* gene led to the loss of resistance against all the diseases mentioned above as well as loss of leaf tip necrosis, which confirmed that *Lr34* is the same gene as *Yr18*, *Pm38* and *Sr37* and *Ltn1* (Krattinger et al., 2009). *Lr34* is an economically important gene that provides partial resistance against all the tested stripe and leaf rust pathogen races and can be referred as 'broad-spectrum' resistance gene.

Two additional non-NLR genes, *Lr46* and *Lr67,* with similar characteristics as *Lr34* were recently described and they both confer APR against leaf, stem, stripe rust and powdery mildew. *Lr67* was cloned by Moore et al. (2015) and based on the mutant screening, the *Lr67* gene was found to encode a member of the hexose sugar transporter family. Like *Lr34*, *Lr67* also confers multiple disease resistance (Moore et al., 2015). Another leaf rust gene which confers APR is *Lr68* but there are no reports so far showing multiple disease resistance and the gene has not yet been identified. APR genes such as *Lr34* and *Lr67* with multiple pathogen resistance are a valuable resource in disease resistance breeding due to their broad effectiveness and durability. Another interesting example in wheat is the *Yr36* gene, which is a partial, broad-spectrum stripe rust resistance gene. This gene encodes for a WHEAT

KINASE START1 (WKS1) protein with a kinase and START domain (StAR-related lipid-transfer) and was originally identified in wild tetraploid emmer wheat (Fu et al., 2009; Uauy et al., 2005).

1.4.4 Other Non-NLR genes for disease resistance in cereals

Also, recently another non-NLR resistance gene *Stb6* was cloned from the old wheat landrace Chinese Spring which provides resistance to *Zymoseptoria tritici*. *Stb6* encodes for a wall-associated receptor kinase (WAK)-like protein which demonstrate disease resistance without HR (Saintenac et al., 2018). There are various examples of resistance genes in cereals that code for wall-associated receptor-like kinases (WAKs). One of the most intriguing example is *Xa4*, a bacterial blight resistance genes in rice that encodes for a WAK (Hu et al., 2017). This single gene is responsible for improving multiple agronomic trait apart from providing resistance in rice such as strengthening of the cell wall and increased lodging resistance and as a result is widely used in rice breeding. In maize, two WAK resistance genes, *Htn1* and *qHSR1* have been identified. *Htn1* confers partial and broad-spectrum resistance against a fungal disease northern corn leaf blight (Hurni et al., 2015). *qHSR1* is a quantitative resistance gene that provides resistance against head smut (Zuo et al., 2015). *qHSR1* is mainly expressed in the mesocotyl and shows an amazing resistance mechanism, where it allows root penetration of the fungus in the plant but later represses the spread of the fungus to the above plant parts. These independent studies highlight the importance of WAKs for disease resistance. Cereal genomes contain hundreds of WAK-like genes. Examples of other cloned QR genes in cereals include *ZmTrxh* which encodes for an atypical thioredoxin and provides resistance to sugarcane mosaic virus in maize (Liu et al., 2017) and *Mlo* gene in barley which encodes for calmodulin-binding protein with seven-transmembrane domain protein and provides resistance to powdery mildew (Buschges et al., 1997), *STV11* which encodes for sulfotransferase and confers resistance to rice stripe virus (Wang et al., 2014) and

Fhb1 which proves resistance to fusarium head blight and encodes for a pore-forming toxin-like domain (Rawat et al., 2016).

1.4.5 Resistance gene deployment strategies

Breeding for rust resistance involves identification of wheat lines with strong resistance to one or more pathogen races. One of the most effective approaches used for resistance breeding is 'gene pyramiding' or stacking. This process involves stacking of several major and minor resistance genes where each gene in the stack is effective against one or several pathogen races of a particular disease to enhance the durability of resistance. This approach of 'gene pyramiding' or stacking is a promising long-term strategy where pathogen mutation is the source of virulence. This means that multiple independent mutations are required in *different* Avr genes for the evolution of virulence in the pathogen. Several APR genes have been effectively used in the breeding program by the CIMMYT breeders using the 'single backcross approach'. This approach is designed to stack multiple, additively acting APR genes with 'minor' effecst in a single genotype to produce a variety or a breeding line with 'near immunity' to rust diseases without the use of the NLR genes effective against specific races of the pathogen in the screening (Singh et al., 2014). Two of the resistance genes with minor effect that have been extensively used in breeding programs are *Lr34* and *Lr46*.

One of the challenges in breeding for durable rust resistance lies in the identification of the best combination of genes to be stacked, which necessitates the cloning of the resistance genes to make effective choices. Stacks of cloned genes can be transferred by transgenesis and such cassettes will be inherited as single genetic unit and are more stable than conventional stacking where unlinked *R* genes can segregate in subsequent generations. Cloning of genes will allow us to characterize functional polymorphisms, which can for example be used to design SNP arrays with all these functional polymorphisms. Moreover, cloning and isolation of gene allows detailed analysis of the molecular mechanisms of resistance, plant-pathogen interactions and

provides a potential novel resistance source which can be altered by genetic engineering for improved function.

1.5 Map-based cloning in wheat – Traditional approach

The most frequently used approach for the cloning of wheat genes has been map-based cloning, also referred to as positional cloning. It is a method to isolate target genes that does not require prior knowledge of the gene product. However, in wheat, map-based cloning is challenging because of the large genome size of 15.4-15.8 Gb, which is five times larger than the human genome and 40 times larger than the genome of rice. More than 85% of the wheat genome is made up of highly repetitive sequences (Wicker et al., 2018). Despite these problems, map-based cloning was the most frequently used approach in wheat for gene cloning, until the recent development of novel gene cloning approaches such as MutRenSeq, MutChromSeq and TACCA, which will be discussed in chapter 4.

The first step in map-based cloning is the development of a bi-parental mapping population (F2, RILs, DH) from two cultivars which differ for the trait of interest. The second step is the development of high-density genetic maps using a combination of genetic and phenotypic data. For this, firstly a low resolution genetic map is established on a small population (100-200) using hundreds of markers. The resolution of this small population is usually around 1-5 cM (Krattinger et al., 2007). To construct a high-resolution genetic map, thousands of plants are then screened with the closest flanking markers and the recombinants are selected for further saturation with molecular markers and phenotyping. This helps in defining a specific chromosomal locus in the genome that carries the gene of interest. For example, *Sr35* was mapped to a 0.98 cM target interval by screening a fine mapping population of 1,925 F2 and 725 BC_1F_1 plants (Saintenac et al., 2013).

The third and the critical step in map-based cloning is the transition from genetic map

to physical map (e.g. from cM to Mb) and to define the physical region spanning the target gene. This is usually done by repeated rounds of bacterial artificial chromosome (BAC) library screening using the closest flanking markers. BAC clones usually have an insert size of 100-200 kb.

The final step in map-based cloning is the validation of the candidate genes. In most of the cases, more than one candidate gene is identified and it is important that these genes are carefully analysed. This is usually done by using mutant screening, virus-induced gene silencing (VIGS), stable transformation or gene knock-out. A widely used approach is by screening for induced loss-of-function mutants. Chemicals such as ethyl methanesulfonate (EMS) or radiation induce random mutations throughout a genome, which will affect the target gene in rare cases. The identification of these loss-of-function mutants and the sequencing of candidate genes can be exploited to validate candidate genes.

Map-based cloning may seem as a simple procedure to follow and isolate the desired gene. But it is important to know and understand the limitations of this time-consuming process in wheat. First, it is very important to develop an accurate mapping population that segregates only for the desired gene of interest. The presence of additional resistance genes in the background will influence the phenotypic evaluation of the population. Secondly, BAC libraries often have uneven coverage of the genomes. For example, a 6x coverage of a genome is usually required to provide a 98% likelihood that all regions are covered (Krattinger et al., 2007). A 6x coverage of a hexaploid wheat genome corresponds to more than 1 million BAC clones. Also, certain genomic regions might not be accessible for cloning into a vector, which will result in gaps in the BAC library. Designing of specific probes to isolate BAC clones spanning the target region is a tedious process due to the high repeat content of the wheat genome (Wicker et al., 2018). Hence a probe which was designed on a repeat sequence will lead to the isolation of a large numbers of BAC clones from off-target

regions. For example, during the cloning of *Lr10*, one repetitive probe was used to screen the BAC library, which identified more than 100 BAC clones (Stein et al., 2000). Another limitation is the hexaploid nature of wheat with three closely related homeologous subgenomes. To overcome this, it is very important that the identified BAC clone is carefully analysed to check if it is from the target genome and target chromosome. Furthermore, ratios of genetic to physical distances are highly variable across the genome and as a result it is difficult to estimate the physical distance and the number of BAC clones that would be required to cover the region between the two flanking markers (Krattinger et al., 2007). For example, during the cloning of *Lr10*, a variable range of recombination frequencies were observed, ranging from 0.6 Mb/cM to 12 Mb/cM in a 230 kb region (Stein et al., 2000).

It is desirable to use a BAC library of a cultivar carrying the gene of interest because of the high diversity and disruption in gene collinearity between different wheat cultivars (Mago et al., 2014). But, due to the high cost and number of clones required to sufficiently cover a wheat genome, it was not feasible to develop BAC libraries for every genotype of interest (Keller et al., 2005).

Despite all these challenges, most of the disease resistance genes such as *Lr1*, *Lr10*, *Lr21*, *Lr34* and *Lr67* in wheat were cloned using map-based cloning. However, *Lr22a* was cloned using a novel approach called 'TACCA' cloning described in Chapter 2. The following paragraphs will highlight the technical advances in wheat genomics which allowed rapid isolation of genes from wheat.

1.6 Advances in wheat genomics to facilitate map-based cloning

More wheat disease resistance genes have been cloned in the past three years than in the 20 years before that. This was possible because of the recent technical advancement in the field of wheat genetics and genomics. Since the advent of the next-generation sequencing (NGS) technologies there has been an upsurge in the speed and effectiveness in the way

genes are mapped and cloned in the hexaploid wheat (Trick et al., 2012). Previously, most of the progress in the wheat genetics and genomics lagged behind other plant species due to the large genome size and hexaploid nature which posed an economic barrier in the whole-genome sequencing of wheat compared to other crop plants such as rice, maize and barley. Therefore, many efforts focused on the development of reduced-representation methods, which targeted specific sequences and allowed generation of molecular markers (Uauy, 2017). For example, SNP genotyping arrays such as the 9k (allowing up to 9,000 markers) (Cavanagh et al., 2013) and 90K Illumina iSelect platforms (allowing 90,000 markers) (Wang et al., 2014) were developed from different wheat accessions of diverse geographical origin for the genotyping of large populations and to generate high-resolution genetic maps.

However, SNP arrays alone did not solve the problem of gene cloning in wheat, therefore other technologies which reduce the complexity of the wheat genome such as RNA-seq, exome capture and chromosome flow-sorting were developed. RNA-Seq (RNA sequencing), also referred as whole transcriptome shotgun sequencing (Wang et al., 2009) is based on the sequencing of the transcribed portion of the genome. In hexaploid wheat, RNA-Seq can reduce genome complexity by approximately 50% (Wulff & Moscou, 2014). RNA-Seq can reveal precise location of the transcription boundaries, to a single-base resolution (Wang et al., 2009) and has been used to identify SNPs to reveal the genetic diversity, to develop SNP based markers for the mapping and for the quantification of the transcriptome. RNA-Seq combined with bulked segregant analysis (BSA) has been used to fine map the stripe rust gene *Yr15* in hexaploid wheat (Ramirez-Gonzalez et al., 2015). Moreover, in wheat, the exome is constituted by 1-2% of the total genome size of which a small fraction is comprised of resistance genes encoding the NLRs (Wulff & Moscou, 2014). Exome capture was used to map *Yr6* locus that is associated with the stripe rust resistance in wheat (Gardiner et al., 2016) and also to design capture array for the NLR complement of potato (Jupe et al., 2013), a technique referred to as Resistance gene enrichment

Sequencing (RenSeq). This method was used on the sequenced potato genome where it allowed identification of additional NLR genes increasing the total number from 438 to 755 (Jupe et al., 2013). The most widely used method for complexity reduction is chromosome flow-sorting. This process involves the isolation of individual chromosome and chromosomal arms using flow cytometry. The basic principle of chromosome flow-sorting is that a macroscopic particle of the sample is passed through the narrow jet which breaks the sample into small droplets. These small droplets carrying the chromosome of interest are then charged electrically and deflected during passage through an electrostatic field (Dolezel et al., 2012) (Vrana et al., 2015). This separates the individual chromosome and chromosomal arms based on their electric charge. However, the utility of these flow-sorted chromosomes depends on the purity and quality of DNA. This is determined using the genomic in situ hybridization, fluorescence in situ hybridisation and G banding. The availability of the individual flow-sorted chromosomes has increased the efficiency and reduced the cost of the sequencing projects. This approach of chromosome flow sorting is nowadays used for many gene cloning projects in wheat.

1.7 Wheat genome sequencing

The major advantage of NGS technologies in wheat research has been in the generation of draft sequences of the hexaploid wheat genome and its diploid progenitors. The presence of three highly similar sub-genomes which diverged 2.5-6.0 million years ago (Chantret et al., 2005) made it difficult to distinguish the sequences between the three sub genomes. Therefore, an alternative strategy of sequencing the diploid progenitors was undertaken which resulted in the sequencing of the *T. urartu* and *Aegilops tauschii*, the donors of the A and D genomes, respectively (Ling et al., 2018; Luo et al., 2017).

The initial draft sequence assemblies of the hexaploid wheat involved a complexity reduction step by isolating the individual chromosomes or chromosomal arms, the process

called as 'chromosome flow-sorting' (Dolezel et al., 2012; Vrana et al., 2015). Using this, *de novo* assembly of the low copy and unique regions was attempted for the flow-sorted chromosome arm 7DS, of the hexaploid wheat landrace, Chinese Spring (Berkman et al., 2011). This same approach was also used to delimit the position of translocation between 7BS and 4AL and reported translocation of approximately 13% of genes from 7BS to 4AL (Berkman et al., 2012). This approach of chromosome flow-sorting was later extended to generate the draft sequences for all the chromosomal arms of wheat except for the chromosome 3B by the International Wheat Genome Sequencing Consortium (IWGSC) (International Wheat Genome Sequencing Consortium, 2014). Chromosome 3B was isolated and used to generate the first high-quality assembly of a wheat chromosome using BAC-by-BAC sequencing (Choulet et al., 2014). The assembly of the 3B provided first insight into the structural and functional portioning of the chromosome. The reference genome sequence of individual chromosomes can be accessed at the IWGSC sequence Repository webpage (http://wheaturgi.versailles.inra.fr/Seq-Repository).

In 2017, a high-quality assembly of the 10.1 Gb wild emmer (*T. turgidum* ssp. *dicoccoides*) was published (Avni et al., 2017) and later in the same year, another high-quality genome sequence of the progenitor of the wheat D genome, *Aegilops tausc*hii was published (Luo et al., 2017). Both of the assemblies were produced using the improved assembly algorithms from the NRGene and provide a detailed insight into the gene content and genome architecture. However, the release of the complete genome sequence of the hexaploid wheat, Chinese Spring (IWGSC RefSeq v1.0) later this year will be the benchmark in the field of wheat research. This availability of this resource will not only provide an understanding of the wheat biology but will accelerate the genome-assisted improvement of the modern wheat varieties.

1.8 Molecular mechanisms of genomic changes

The grasses (in particular Brachypodium, barley, maize, sorghum, rice and wheat) have served as model plant family for comparative genetics and genomics for the last two decades. Grasses are all derived from a single common ancestor that lived 50-80 million years ago. Despite the recent and monophyletic origin, the grass species have diverged tremendously with respect to chromosome number and genome size (Bennetzen, 2007). Previously, one of the vital tools in grass comparative genomics has been the collinearity of the genetic maps and this was first evidenced by intraspecies recombinational maps which were based on shared DNA markers (Bennetzen, 2007). The collinearity of the genes was later confirmed by DNA sequence analysis of small chromosomal segments from orthologous regions. Comparative studies revealed that most of the gene positions were retained for most grass loci but numerous small genic rearrangements by genomic DNA insertions were observed for which the mechanisms remained unclear. In some cases, gene loss was observed which resulted from small deletions or gene inversions by unequal cross over events between flanking repeats. Common causes of insertions are transposable elements (Ma & Bennetzen, 2004; Wicker et al., 2016), unequal recombination (Woodhouse et al., 2010) and ectopic recombination stimulated by double-strand breaks (DSBs) (Salomon & Puchta, 1998; Wicker et al., 2010).

1.8.1 Genomic changes mediated by DNA transposons

Grass genomes contains enormous number of DNA transposons. DNA transposons move in the genome by excising from or by inserting into the genomic DNA. The excision of the DNA transposons causes DSBs which have to be repaired by the cell. Previous studies have shown that these excisions and insertions of the DNA transposon can lead to deletions and insertions of filler sequences (depending on the repair pathway) at the site of the DSBs (Roffler et al., 2015; Roffler & Wicker, 2015). However, sometimes these re-arrangements at the excision site can be so extensive, that it is difficult to identify the excision site. Wicker et

al (2016) proposed a model for increased mutation rates for the genes of rice which was caused by transposon excision and insertion (Fig. 1.4). In the first step, transposons excise from the genome which causes a DSB for a cell to repair. After excision, exonucleases produce 3' overhangs which are annealed using micro-homologies of few base pairs. The single-stranded DNA segments are used as template for the synthesis of new second strand which introduces numerous mutation (Wicker et al., 2016).

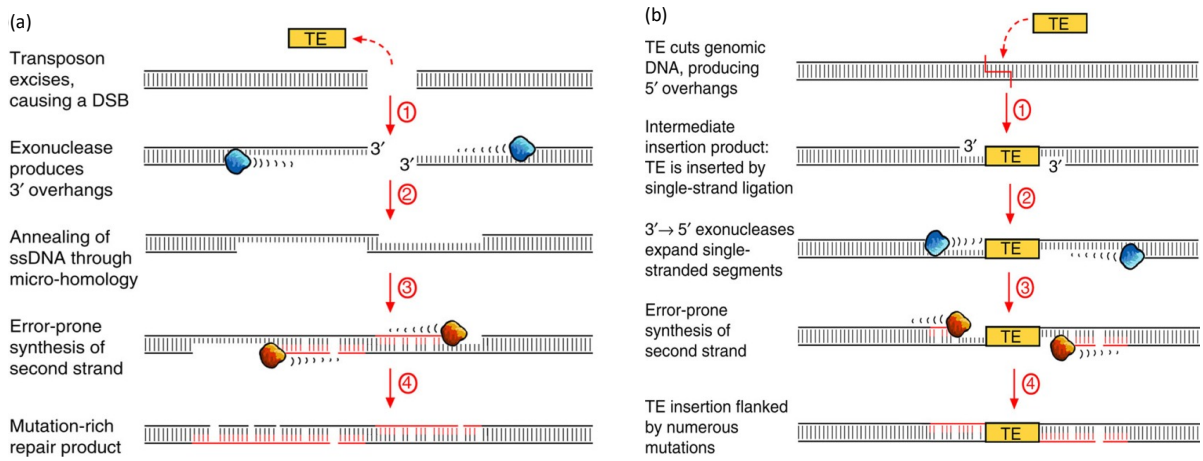


Fig. 1.4 Genomic changes caused by excision (a) and insertion (b) of DNA transposon into the genomic DNA. (Source: Wicker et al., 2016).

1.8.2 Genomic changes mediated by mechanisms other than DNA transposons

DSB occur frequently at the fragile sites which consist of tandem repeats motifs such as micro- and minisatellites (Wicker et al., 2010). These tandem repeats are hotspot for recombination by unequal cross-over (see chapter 3) or template slippage. Duplicated regions flanked by tandem repeats on both sides have been described in rice and *Brachypodium* (Fig. 1.5). In rice, it was shown that the duplicated fragment that contained rice gene *Os3g30240*, was located inside an array of tandem repeats. There were three units on the left side and five repeats units on the right side, both were GC rich.

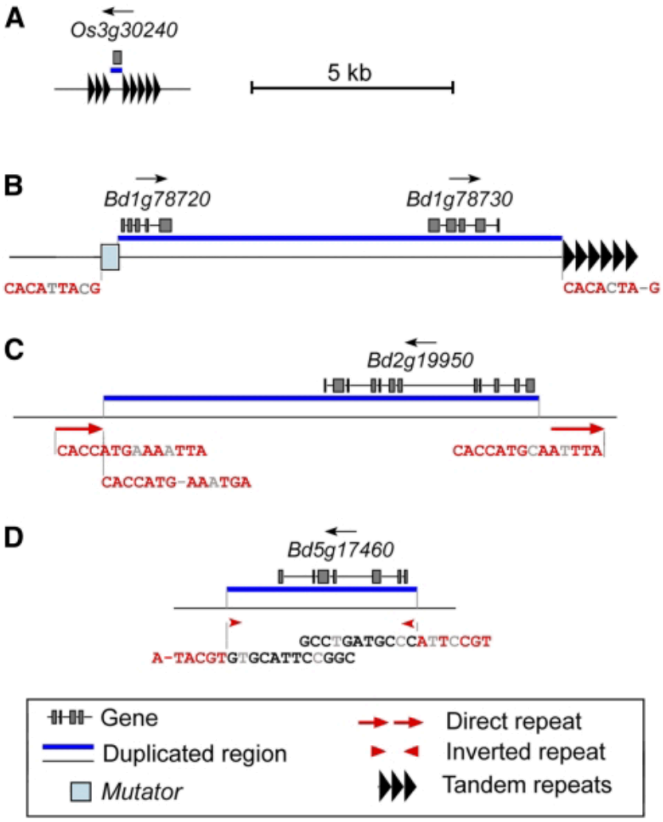


Fig. 1.5 Molecular mechanisms of genomic changes caused by mechanisms other than transposable elements. (Source: Wicker et al., 2010).

It was assumed that during the template slippage or unequal cross-over, DSB occurred which was then repaired with foreign fragment containing the gene. Also, in *Brachypodium*, it was observed that the duplicated fragment which contained *Bradi78720* and *Bradi1g78230* was flanked on one side by *Mutator* element and on other side by large array of direct repeats. It was hypothesised that the *Mutator* element caused the DSB in the unstable region.

1.9 Aim of the thesis

The aim of the thesis was:

(i) To develop a novel technology for the rapid isolation of disease resistance genes (*Lr22a*) in wheat using the high-quality sequence from the parent of interest ('CH Campala *Lr22a*').

(ii) To use this high-quality sequence of 'CH Campala *Lr22a*' for the comparative analysis with the high-quality sequence of 'Chinese Spring' (IWGSC RefSeq v1.0) to identify genomic differences between the two wheat cultivars.

Chapter 2

Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly

Anupriya Kaur Thind[+], Thomas Wicker[+], Hana Šimková, Dario Fossati, Odile Moullet, Cécile Brabant, Jan Vrána, Jaroslav Doležel, Simon G. Krattinger

[+]These authors contributed equally to this work

(2017)

Nature Biotechnology 5(8):793-796.

doi: 10.1038/nbt.3877

Abstract

Cereal crops such as wheat and maize have large repeat-rich genomes that make cloning of individual genes challenging. Moreover, gene order and gene sequences often differ substantially between cultivars of the same crop species (Chia et al., 2012; Jordan et al., 2015; Mago et al., 2014; Rawat et al., 2016). A major bottleneck for gene cloning in cereals is the generation of high-quality sequence information from the cultivar of interest. In order to accelerate gene cloning from any cropping line, we report 'targeted chromosome-based cloning via long-range assembly' (TACCA). TACCA combines lossless genome complexity reduction via chromosome flow sorting with Chicago long-range linkage (Putnam et al., 2016) to assemble complex genomes. We applied TACCA to produce a high-quality (N50 of 9.76 Mb) *de novo* chromosome assembly of the wheat line 'CH Campala *Lr22a*' in only four months. Using this assembly, we cloned the broad-spectrum *Lr22a* leaf-rust resistance gene using molecular marker information and ethyl methanesulfonate (EMS) mutants, and found that *Lr22a* encodes an intracellular immune receptor homologous to the *Arabidopsis thaliana* RPM1 protein.

2.1 Introduction, result and discussion

While the world population continues to grow, the arable land per capita is decreasing (FAO). To ensure food security, agriculture will require high-yielding crops that can withstand diseases, pests and adverse climatic conditions. A better understanding of the genes that control these important traits may enable breeding of crop cultivars capable of feeding the 9–10 billion people expected to be living by 2050.

'Positional cloning' or 'map-based cloning' is often used to clone plant genes (Krattinger et al., 2007). Unlike other gene cloning strategies, positional cloning requires no prior knowledge of the gene sequence or product. One crucial step during positional cloning is the production of high-quality genome sequence information spanning the region that contains the gene of interest. Although a reference genome sequence can serve as a 'guide' to narrow down the location of a gene, the gene causing the phenotype of interest is often absent from the reference cultivar (Mago et al., 2014; Rawat et al., 2016), which means that sequence information from a line that carries the gene of interest is needed. Repeated rounds of bacterial artificial chromosome (BAC) library screening, or chromosome walking, are usually necessary to cover the region of interest with a contiguous sequence (Stein et al., 2000).

Chromosome walking is particularly tedious in crop species that have large and repeat-rich genomes, such as wheat. The main limitation of BAC clones is that they can only harbor inserts of ~100–200 kb, which is why chromosome walking can take a long time. Sequencing and assembly technologies that produce longer sequence scaffolds could prove to be particularly advantageous for positional cloning of plant genes. It has recently been shown that chromosome conformation capture technologies provide powerful tools that enable the assembly of short sequence reads into long, megabase-sized scaffolds in humans and *Drosophila melanogaster* (Ay & Noble, 2015; Burton et al., 2013).

Leaf rust, caused by the pathogenic fungus *Puccinia triticina*, is a widespread and devastating disease of wheat (Kolmer, 2013) that can be sustainably controlled by exploiting disease resistance that is present in some cultivars of this crop. The disease resistance gene *Lr22a* was crossed into hexaploid bread wheat (*Triticum aestivum*) from its wild relative *Aegilops tauschii* in the 1960s (Dyck & Kerber, 1970). Following this initial cross, *Lr22a* has subsequently been bred into several Canadian wheat cultivars (Hiebert et al., 2007), and *Lr22a*-containing wheat lines have been included in leaf rust surveys worldwide for many years. *Lr22a* confers resistance to a wide range of *P. triticina* isolates (Hiebert et al., 2007; Kolmer, 1997; McCallum et al., 2013; Pretorius et al., 1987). The *Lr22a*-mediated resistance is not present in young seedlings (<20 d) but is only visible in wheat plants from ~25 d of age.

First, in order to evaluate the effectiveness of *Lr22a* against Swiss *P. triticina* isolates, we inoculated the *Lr22a*-containing backcross line RL6044 (Thatcher *Lr22a*) and the spring wheat cultivar Thatcher with ten *P. triticina* isolates that were collected in Switzerland. The first leaves of RL6044 developed leaf rust pustules of similar size as those of the susceptible control Thatcher, while we observed complete to moderate resistance on the third leaves of 30-d-old RL6044 plants in comparison to Thatcher (Fig. 2.1 and Supplementary figure S2.1).

Lr22a was previously mapped to the short arm of wheat chromosome 2D using microsatellite analysis of the *Lr22a*-containing wheat line 98B34-T4B13. In order to pinpoint *Lr22a* on chromosome 2D we generated a high-resolution mapping population from a cross between the susceptible Swiss spring wheat cultivar CH Campala and an *Lr22a*-containing backcross line CH Campala *Lr22a* (Moullet et al., 2014) and delimited the gene to a 0.48-cM interval flanked by two microsatellite markers gwm455 and gwm296 (Fig. 2.2a).

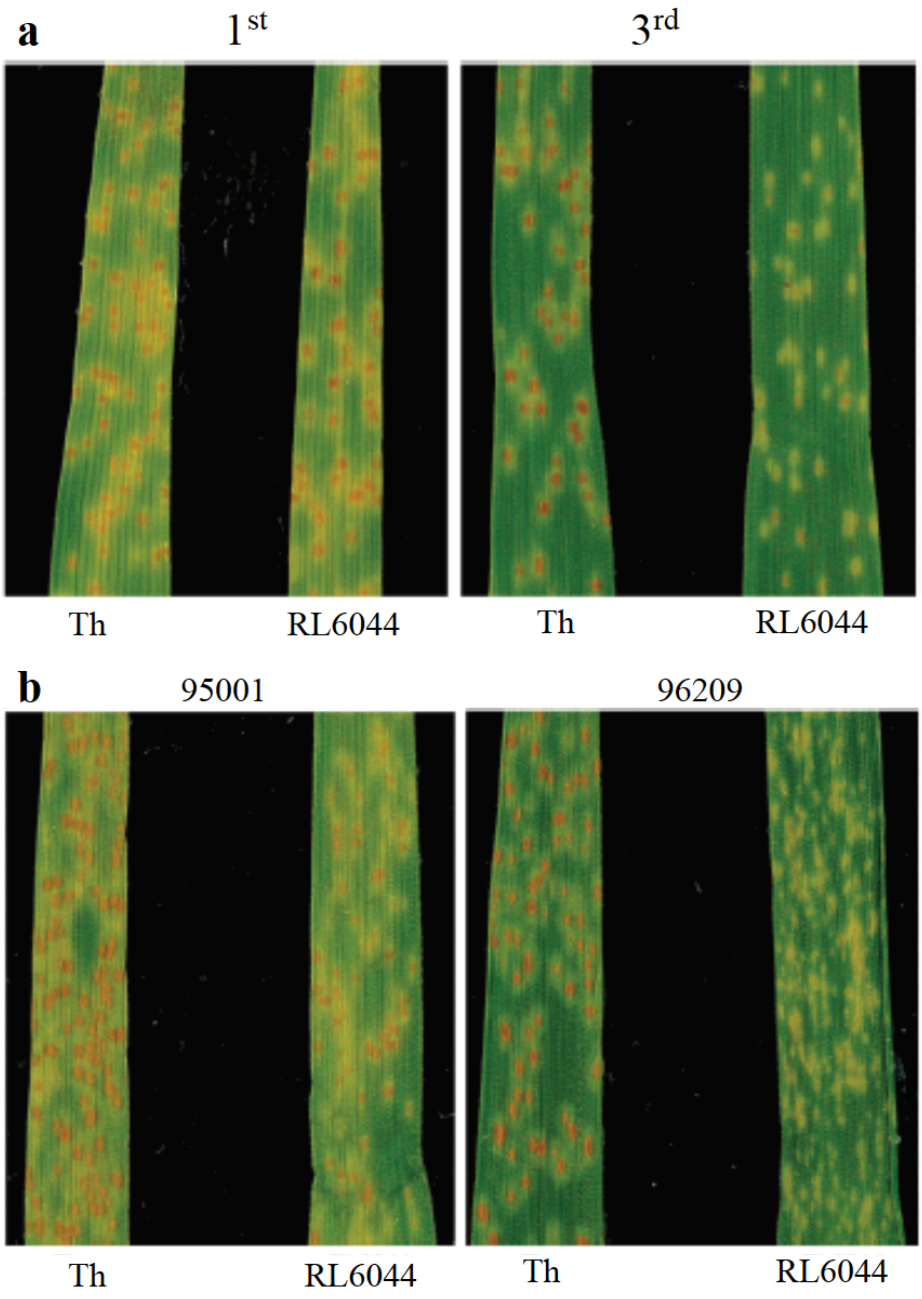


Fig. 2.1 Phenotypic response conferred by the *Lr22a* leaf rust resistance gene. (**a**) Leaf rust symptoms on first and third leaves of 30-d-old plants of the susceptible cultivar Thatcher (Th) and the *Lr22a*-containing backcross line RL6044 (Thatcher *Lr22a*). (**b**) The *Lr22a*-resistance response in RL6044 ranged from partial (left) to complete (right) against different *P. triticina* isolates. Shown here are the two extremes found with *P. triticina* isolates 95001 and 96209.

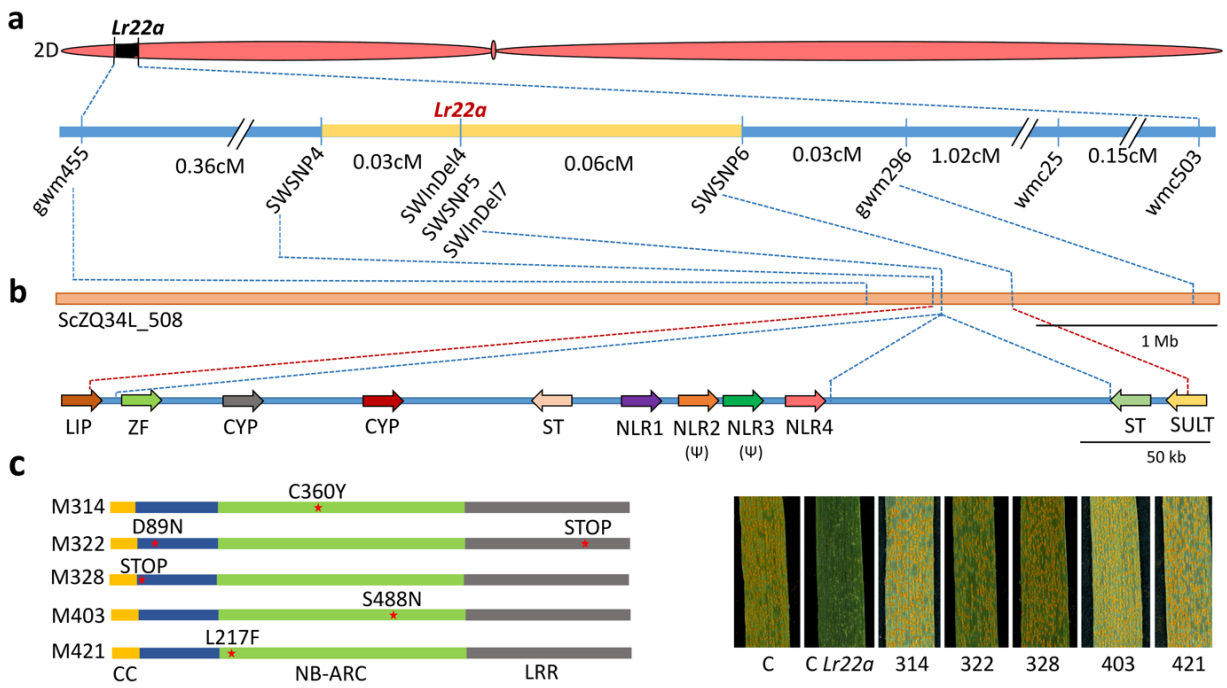


Fig. 2.2 Mapping of the *Lr22a* leaf rust resistance gene. (**a**) Genetic map of the *Lr22a* region. The target interval between the closest flanking markers SWSNP4 and SWSNP6 is indicated in yellow. (**b**) The physical interval of CH Campala *Lr22a* contained nine candidate genes and two pseudogenes (indicated by Ψ). LIP = lipase, ZF = zinc finger, CYP = cytochrome P450, ST = sugar transporter, NLR = nucleotide binding site–leucine-rich repeat receptor, SULT = sulfotransferase. The 6.39-Mb sequence scaffold ScZQ34L_508 that contained both flanking markers is indicated in orange. (**c**) Five independent EMS mutants that lost the *Lr22a*-resistance response had non-synonymous sequence changes in the *NLR1* coding sequence compared to the wild-type allele of CH Campala *Lr22a*. The predicted coiled-coil (CC), nucleotide-binding (NB-ARC) and leucine-rich repeat (LRR) domains of the NLR1 protein are indicated in yellow, green and gray, respectively. The amino acid polymorphisms in comparison to the Lr22a wild-type sequence of CH Campala *Lr22a* are indicated by red asterisks. C: CH Campala; C *Lr22a*: CH Campala *Lr22a*.

Then, to obtain sequence information for this 0.48-cM interval, we isolated chromosome 2D from CH Campala *Lr22a* using flow cytometry (Dolezel et al., 2012), and obtained ~640 ng high-molecular-weight DNA of this chromosome. On the basis of recent examples of long-range scaffolding with the help of chromosome contact maps (Ay & Noble, 2015; Burton et al., 2013), we obtained a *de novo* assembly of chromosome 2D by combining short-read Illumina sequences and proximity ligation of *in vitro*–reconstituted chromatin, also known as Chicago (Putnam et al., 2016). In contrast to the *in vivo* Hi-C method, Chicago has been demonstrated to be more suitable for generating high-quality assemblies from short Illumina contigs in vertebrates (Putnam et al., 2016).

The assembly of CH Campala *Lr22a* comprised 10,344 scaffolds with an N50 of 9.76 Mb, that is, half of the chromosome was assembled in scaffolds of 9.76 Mb or more. This N50 is 50–100× longer than a BAC clone, and thus each scaffold of this length corresponds to at least 25 rounds of BAC library screening. The longest scaffold was 36.4 Mb and the total assembly was 567 Mb (Supplementary table S2.1). The size of the assembly was ~160 Mb shorter than the estimated size of chromosome 2D (Safar et al., 2010), which was likely due to collapsed high-copy repeats in the assembly (Supplementary figure S2.2). The flanking markers gwm455 and gwm296 were located at a distance of 1.79 Mb on a single scaffold (ScZQ34L_508) of 6.39 Mb in size (Fig. 2.2b). We used this CH Campala *Lr22a* scaffold to develop additional markers by comparing annotated gene sequences to Illumina reads of the susceptible parent line CH Campala. This allowed us to further reduce the genetic interval to only 0.09 cM (Fig. 2.2a). The physical distance between the two flanking markers SWSNP4 and SWSNP6 was 438 kb and contained nine genes and two pseudogenes (Fig. 2.2b). In particular, there was a cluster of two genes encoding nucleotide binding site–leucine-rich repeat receptor (NLR) and two NLR pseudogenes. *NLR1* showed sequence alterations compared to the wild-type CH Campala *Lr22a* allele in five independent EMS mutants that were generated from CH Campala *Lr22a* and

that were identified using a phenotypic screen for loss of *Lr22a* resistance. All of the single-nucleotide polymorphisms (SNPs) present in the susceptible mutants are predicted to result in amino acid exchanges or premature stop codons in NLR1 (Fig. 2.2c and Supplementary table S2.2). The sequence of the second full-length *NLR4* gene was identical to the wild-type sequence of CH Campala *Lr22a* in all five susceptible mutants. These results provide evidence that *NLR1* corresponds to the *Lr22a* resistance gene.

To evaluate the overall quality of the CH Campala *Lr22a* assembly, we anchored the scaffolds to a high-resolution genetic map of the wheat D-genome progenitor *Ae. tauschii* that includes 1,326 chromosome-2D-specific SNP markers (Luo et al., 2013). To do this, we performed a BLAST search (Altschul et al., 1997) with the extended sequences of the *Ae. tauschii* SNP markers against the CH Campala *Lr22a* assembly. In total, 1,048 sequences produced BLAST hits that anchored 80 scaffolds (or 521 Mb, which is 92% of the assembly) to the genetic map (Supplementary table S2.3). Each of the anchored scaffolds contained an average of 13 SNP markers (ranging from 1–83 markers). We observed a high degree of collinearity between *Ae. tauschii* and the CH Campala *Lr22a* assembly (Supplementary figure S2.3a). Only 62 of the 1,048 genetic markers were non-collinear (mapped to a different location than most markers on the scaffold). Of these, 44 markers were grouped into seven clusters, meaning that at least two markers mapped to a different region on the *Ae. tauschii* genetic map than most of the markers on the scaffold. This might indicate the presence of seven chimeric scaffolds in which two large genomic segments were incorrectly joined. Alternatively, the non-collinear markers might arise by structural variation. The remaining 18 non-collinear markers represented single markers that mapped to a different *Ae. tauschii* position than all others on the respective scaffold; this might be explained by problems in the genetic map of *Ae. tauschii*. SNP markers were perfectly collinear within the *Lr22a* region (Supplementary figure S2.3b).

The predicted coding sequence of *Lr22a* is 2,739 bp, consists of a single exon, and translates into a protein of 912 amino acids with an N-terminal coiled-coil domain, a central nucleotide-binding (NB-ARC) domain, and a C-terminal leucine-rich repeat domain. In the susceptible parent CH Campala, the *NLR1* allele was disrupted by a premature stop codon, whereas the *NLR1* allele in the susceptible wheat line Thatcher was complete and the predicted protein showed 97% amino acid identity to Lr22a (Supplementary figure S2.4). The Lr22a protein showed only weak sequence homology to other cloned wheat NLRs. The closest homolog of Lr22a in *Arabidopsis* is RPM1, an NLR that confers resistance to the bacterial pathogen *Pseudomonas syringae* (Supplementary figure S2.5). The N-terminal amino acids of RPM1 interact with the RPM1-interacting protein 4 (RIN4), an important regulator of basal defense responses that is targeted by multiple *P. syringae* virulence effectors. Effector-mediated modification of RIN4 is perceived by RPM1, resulting in a hypersensitive response (Belkhadir et al., 2004; Mackey et al., 2002). Similarly, it is possible that Lr22a might monitor the status of a basal defense component in wheat. Interestingly, Lr22a contains two amino acids at the N terminus that are unique compared to the NLR1 protein variants in 25 wheat cultivars without the *Lr22a* resistance (Supplementary figure S2.6).

Several rapid gene cloning methods have been described for wheat (Choulet et al., 2014; Gardiner et al., 2016; Sanchez-Martin et al., 2016; Steuernagel et al., 2016) (Table 2.1). All of these approaches require the identification of loss-of-function mutants, and some of the methods, such as MutRenSeq, are only suitable for specific gene classes. However, many agriculturally important genes, for example, genes conferring partial disease resistance or abiotic stress tolerance, have 'partial phenotypes' for which the identification of loss-of-

Table 2.1. Comparison of different gene isolation approaches

	MutChromSeq	MutRenSeq	Mapping-by-sequencing	Positional cloning by chromosome walking	Targeted-chromosome-based-cloning via long-range assembly
Dependence on enrichment library	No	Yes	Yes	No	No
Dependence on the identification of loss-of-function mutants	Yes	Yes	Yes	No	No
Dependence on reference sequence	No	Depends on reference gene annotation for enrichment library	Depends on reference gene annotation for enrichment library	No	No
Speed / cost-effectiveness	Very rapid, cost-effective	Very rapid, cost-effective	Very rapid, cost-effective	Very slow, expensive	Rapid, cost-effective
Major limitations	Depends on the identification of loss-of-function mutants, no backup if mutants cannot be identified	Only allows identification of NLRs, depends on enrichment library, depends on the identification of loss-of-function mutants, no backup if gene of interest does not encode a NLR	Depends on enrichment library or a high-quality reference sequence, depends on the identification of loss-of-function mutants	Very slow, a cultivar-specific BAC library is often necessary, depends on recombination	Partially depends on recombination, but also works in chromosomal regions with reduced recombination rates
Best suited for	Isolation of genes with strong phenotypes	Isolation of NLRs with strong phenotypes	Isolation of genes with strong phenotypes	Any gene, also suitable for genes with partial phenotypes and adult plant phenotypes	Any gene, also suitable for genes with partial phenotypes and adult plant phenotypes

function mutants can be challenging. TACCA offers greater flexibility with respect to gene validation (e.g., transformation, haplotype analysis, TILLING, or genome editing), and since it includes the generation of mapping population, this approach enables positional cloning of genes with partial phenotypes.

Using a cultivar-specific *de novo* assembly, we eliminated the need for chromosome walking. Positional cloning requires high-density genetic maps, which are attainable only in distal, telomeric chromosome regions that are characterized by high recombination rates. Pericentromeric and centromeric chromosomal regions show lower recombination rates, which makes the construction of high-density genetic maps challenging.

Long-range scaffolding approaches, however, permit gene cloning even in regions with lower recombination rates, such as pericentromeric regions and alien introgressions. For example, for the telomeric region of chromosome arm 2DS, a mapping population of only 400 plants would have been sufficient to reach a 96% probability of finding a target gene and its closest flanking markers on the same sequence scaffold. In pericentromeric regions, where recombination rates are 5–10× lower, a mapping population of 1,200 plants would provide a 90% chance to find a target gene and its closest flanking markers on a single sequence scaffold (Supplementary figure S2.7).

Gene density in wheat and many other grass genomes is highest in distal regions of the chromosome (Choulet et al., 2014; Gottlieb et al., 2013; Stein et al., 2001). Therefore, our approach could find widespread application in cloning most genes in cereals. Crucial for the long-range scaffolding of the *Lr22a* region was the amount of DNA required for sequencing because this determined the time needed for chromosome purification. Chicago scaffolding works with small amounts of DNA (~500 ng) and was therefore well-suited to enable a high-quality *de novo* assembly from a flow-sorted chromosome. However, other long-range

scaffolding or long-read sequencing technologies that work with small amounts of DNA (<1 µg) such as nanopore sequencing, for example, might also be used in our gene cloning strategy.

In summary, we report that it is now feasible to develop high-quality *de novo* assemblies from chromosomes of any wheat cultivar. Our approach can be applied in species with complex genomes, which should enable cloning of agriculturally important genes. Any species and cultivar from which chromosomes can be flow-sorted can be used. To date, flow cytometry has been successfully used in more than 20 plant species, including important crops like maize, wheat, rice, barley, oat, rye, pea, tomato, field bean, and chickpea (Dolezel et al., 2012).

2.2 Methods

2.2.1 Plant material

The *Lr22a*-containing wheat lines RL6044 (Thatcher*7//tetra-Canthatch/RL5271) and CH Campala *Lr22a*17 (CH Campala*6/AC Minto) were used in this study. A bi-parental mapping population consisting of 1,656 F2 plants was derived from a cross between CH Campala *Lr22a* and the susceptible near-isogenic Swiss spring wheat cultivar CH Campala. DNA was extracted from leaf tissues using a cetrimonium bromide extraction protocol (Stein et al., 2001). A total of 1,656 F2 plants were screened for recombination between simple sequence repeat (SSR) markers gwm455 and wmc503 (Hiebert et al., 2007) PCR products were separated on polyacrylamide gel using a LI-COR DNA Sequencer 4200. In total, 54 recombinant F2 plants were identified showing 55 recombination events between the two markers. F3 families of recombinant F2 plants were phenotyped in the field and growth cabinets. F3 families were classified as uniform susceptible, uniform resistant, or segregating based on a comparison to the two parents. In addition, homozygous recombinant F4 families were selected and re-phenotyped in growth cabinets. Field infections were done as described previously (Singla et al., 2017) or

the infection assays in growth cabinets, five seeds per family were sown in two replicates in soil in 1.5-liter pots. After treatment with 4 ml/l growth inhibitor (Cycocel Extra, Omya AG, Oftringen, Switzerland) and 2–3 ml/l fertilizer (Wuxal Profi, Maag Garden, Syngenta, Dusseldorf, Germany) plants were grown at 20 °C and a 16 h photoperiod (450 µmol m−2 s−1) followed by 8 h at 16 °C without light and a relative humidity of 70%. Plants were inoculated with *P. triticina* isolate 90035 suspended in oil (Fluorinert FC-43, 3M Electronics, Zwijndrecht, Belgium) when they were 20- to 25-d old. After the inoculation, plants were kept in the dark for 24 h under a plastic tent to maintain high humidity and then shifted back to normal growth conditions. Disease symptoms were assessed 10 d after inoculation.

2.2.2 EMS mutagenesis and identification of *Lr22a* mutants

Ethyl methanesulfonate (EMS) mutagenesis was performed as described previously (Periyannan et al., 2013). In a preliminary experiment, 0.35% EMS was identified as the concentration that resulted in 50% seedling mortality. Then, 1,100 seeds of CH Campala *Lr22a* were soaked in water at 4 °C for 16 h and then the treatment with 0.35% EMS was done for 16 h at room temperature with constant shaking at 150 r.p.m. Treated seeds were washed in tap water and plants were advanced to M2 generation in the glasshouse. Seeds of 685 M1 plants were harvested and the respective M2 families were screened for susceptibility with the *P. triticina* isolate 90035 as described above. Out of this screen, five susceptible mutants derived from different M2 families were identified and validated in the M3 generation. All susceptible mutants identified in this screen carried sequence polymorphisms in NLR1.

2.2.3 Flow sorting of chromosome 2D and preparation of DNA samples

Chromosome 2D was purified by flow cytometric sorting as described earlier (Kubalakova et al., 2002; Vrana et al., 2000) with modifications. Briefly, suspensions of intact

mitotic metaphase chromosomes were prepared from synchronized root tip meristem cells. Before flow cytometry, GAA microsatellites were labeled on chromosomes in suspension by FITC following a previously described protocol (Giorgi et al., 2013), and chromosomal DNA was stained by DAPI (4′,6-diamidino 2-phenylindole). Chromosome samples were analyzed at rates of 1,500 –2,000 particles/sec on a BD FACSAria SORP flow cytometer (Becton Dickinson Immunocytometry Systems, San Jose, USA) and bivariate flow karyotypes FITC vs. DAPI fluorescence were acquired. Sort windows delimiting the population of chromosome 2D were set on dotplots fluorescein isothiocyanate (vs. DAPI (Supplementary figure 2.8), and chromosome 2D was sorted at rates of 15–20/sec. The identity of flow-sorted chromosomes and contamination by other chromosomes were checked microscopically using fluorescence *in situ* hybridization (FISH) as described previously (Kubalakova et al., 2003) using probes for GAA microsatellites and *Afa* family repeat.

For shotgun sequencing, DNA of chromosome 2D was amplified by multiple displacement amplification (MDA) as described previously (Simkova et al., 2008). In total, 30,000 copies of chromosome 2D were flow-sorted from each line. The purity of the sorted fraction was 94%. The chromosomes were treated with proteinase K and the purified DNA was amplified using an Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Chalfont St. Giles, UK). Three independent MDA products from each sorted chromosome fraction were pooled into one sample to reduce amplification bias.

For long-range assembly, high molecular weight (HMW) DNA was prepared from flow-sorted chromosome 2D of CH Campala *Lr22a* following a previously described protocol (Šimková et al., 2003) with modifications. A total of 1.5 million copies of chromosome 2D were flow-sorted with a purity of 97% and embedded in six agarose miniplugs with a total volume of 100 µl. Plugs were then incubated in proteinase K. The miniplugs were washed six times in TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0), melted

for 5 min at 73 °C and solubilized with 0.8 U GELase (Epicentre, Madison, USA) for 45 min. The released DNA underwent 60 min of drop dialysis (Merck Millipore, Billerica, USA) against TE buffer. Purification and concentration was performed using a Vivacon 500 centrifugal concentrator (100,000 Dalton MWCO, Sartorius, Goettingen, Germany). The HMW DNA was partially fragmented by pipetting and vortexing to facilitate concentration measurement.

2.2.4 Establishment of long-range assembly from CH Campala *Lr22a*

Chromosome 2D shotgun sequencing, Chicago sequencing and scaffolding was performed by Dovetail Genomics (Santa Cruz, CA). A Chicago library was prepared as described previously (Putnam et al., 2016). Briefly, 250 ng of chromosome 2D HMW DNA (mean fragment length ~100 kb) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with MboI, the 5′ overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350-bp mean fragment size and a sequencing library was generated using NEBNext Ultra enzymes (New England BioLabs) and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment. The library was then sequenced on an Illumina HiSeq 2500 (rapid run mode) to produce 145 million 150-bp paired-end reads, which provided 30× physical coverage of the chromosome (1–50 kb pairs).

De novo chromosome 2D assembly was constructed using sequence data from three paired-end libraries, two prepared from 50 ng of chromosomal DNA with a mean insert size of 205 bp and one prepared from 150 ng of chromosomal DNA with a mean insert size of 450 bp. The libraries were sequenced on an Illumina HiSeq 2500 (rapid run mode) to produce a total of 709 million 150-bp paired-end reads (312 million from the shorter

insert libraries and 397 million from the longer insert library). Reads were trimmed for quality, sequencing adapters, and mate pair adapters using Trimmomatic (Bolger et al., 2014). *De novo* assembly was performed using Meraculous 2 (2.2.2.3) (Chapman et al., 2011) with a k-mer size of 109.

The input *de novo* assembly, shotgun reads, and Chicago library reads were used as input data for HiRise, a software pipeline designed specifically for using Chicago data to scaffold genome assemblies (Putnam et al., 2016). Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify putative misjoins and to score prospective joins. After scaffolding, the shotgun sequences were used to close gaps between contigs. To generate a pseudomolecule, the extended sequences of 1,326 chromosome 2D-specific SNPs mapped to the *Ae. tauschii* AL8/78 genetic map (Luo et al., 2013) were used to perform a BLAST search against the 10,344 CH Campala *Lr22a* scaffolds using an in-house script.

2.2.5 Marker development

SSR markers gwm455, gwm296, wmc503, and wmc25 were previously reported to be linked to *Lr22a* (Hiebert et al., 2007). For the development of additional markers, a *de novo* Illumina sequence assembly was developed from DNA amplified from flow-sorted 2D chromosomes of CH Campala and CH Campala *Lr22a*. DNA from chromosome 2D of each parent were multiplexed and sequenced on one lane of an Illumina HiSeq 2500 with 125-bp paired-end reads. The sequencing was performed at the Functional Genomics Center Zurich, Switzerland. The reads were used for a *de novo* assembly using CLC Main Workbench 7 (Qiagen) with standard parameters and a minimum contig length of 500 bp. For CH Campala, 84 million reads were obtained and assembled into 57,314 contigs with a total size of 123 Mb and a scaffold N50 of 3.8 kb. For CH Campala *Lr22a*, 139 million reads were obtained that

were assembled into 71,348 contigs with a total size of 159 Mb and a scaffold N50 of 3.9 kb. Illumina contigs were filtered for contigs containing genes by performing a BLAST search (Altschul et al., 1997) against the *Brachypodium distachyon* coding sequence database (International Brachypodium Initiative, 2010). Gene-containing contigs were used for the discovery of SNPs and insertions/deletions (InDels) based on a previous protocol (Shatalina et al., 2013) with minor modifications. The sequences of the *Lr22a* flanking SSRs gwm455 and wmc25 were anchored to the genetic map of *Ae. tauschii* AL8/78 (Luo et al., 2013) by performing a BLAST search against the *Ae. tauschii* BAC scaffolds (Jia et al., 2013; Luo et al., 2013) (http://aegilops.wheat.ucdavis.edu/ATGSP/). This resulted in the identification of two scaffolds, 4242.1 (gwm455) and 4531.6 (wmc25). A second BLAST search with the identified scaffolds against the extended sequences of the *Ae. tauschii* SNP markers (http://probes.pw.usda.gov/WheatDMarker/) identified the chromosome 2D-specific markers AT2D1039 and AT2D1040 (gwm455) and AT2D1053 (wmc25) that were located at cM positions 25.59–28.502 on the *Ae. tauschii* genetic map. The extended sequences of markers mapped between AT2D1039 and AT2D1053 were then used to perform a BLAST search against the *Ae. tauschii* BAC scaffolds. The *Ae. tauschii* BAC scaffolds were used to identify the corresponding sequences in the gene-containing contigs of CH Campala and CH Campala *Lr22a*. The identified contigs of the two wheat lines were aligned using Clustal Omega (Sievers et al., 2011), and locus-specific PCR probes spanning polymorphisms between CH Campala and CH Campala *Lr22a* were developed and sequenced on the recombinants of the fine-mapping population. This resulted in the development of two markers, SWSNP5 and SWInDel4 (Supplementary table S2.4). Similarly, the CH Campala *Lr22a* Chicago assembly was used to develop additional markers. Scaffold ScZQ34L_508 was annotated using the *B. distachyon* coding sequence database (https://phytozome.jgi.doe.gov/pz/portal.html). The Illumina contigs of CH Campala were mapped against the annotated genes using BLAST and

SNPs and InDels were identified as described above. This resulted in the development of three additional markers, SWSNP4, SWSNP6, and SWInDel7. For the amplification of *Lr22a*, specific primers (LRR1-F3 and LRR1-R4) were designed from the 5′ and 3′ UTR and amplified using the Kapa HiFi HotStart PCR kit (KapaBiosystems) according to the manufacturer's protocol. The amplicon was sequenced using eight internal primers (Supplementary table S2.4).

2.2.6 Lr22a protein domain prediction

The predicted Lr22a protein sequences from RL6044 and corresponding NLR1 protein version from Thatcher were aligned using the online Clustal Omega (Sievers et al., 2011). Different domains of the NLR were identified based the homology to the annotated RPM1 protein (Gao et al., 2011). The most probable LRR motifs were predicted using the LRR conservation mapping tool v2.0 (http://www.plantpath.wisc.edu/RCM) (Helft et al., 2011).

2.2.7 Statistical methods

A phylogenetic tree of Lr22a and known wheat resistance proteins was made using the PROTPARS tool of the PHYLIP package with 100 bootstrap replicates (Retief, 2000). The amino acid sequences of known wheat NLRs were downloaded from the NCBI repository. Amino acid sequences of the LRRs were aligned using ClustalX 2.1 using a gap opening penalty of 10 and a gap extension penalty of 0.2.

2.2.8 Simulation of recombination frequencies and population sizes

The goal of this simulation was to calculate the probabilities of finding a target gene and its closest flanking markers on a single sequence scaffold using different sizes of mapping populations. Recombination frequencies were derived from combining the genetic mapping data from *Ae. tauschii* (Luo et al., 2013) and the physical sizes of the 80 CH Campala *Lr22a* scaffolds that were anchored to the genetic map (Supplementary table S2.3). Local recombination

frequencies (in Mb/cM) along chromosome 2D were calculated in a sliding window averaging ratios of physical to genetic distance over 50 genetic markers. Based on the resulting recombination frequency graph, we divided the chromosome into two telomeric, two pericentromeric and one centromeric bin (Supplementary figure S2.7a). Simulations were run for the two telomeric bins separately, because recombination frequencies on the short- and long-arm telomeric bins differed by a factor of 2.3 (median 1.2 Mb/cM for 2DS vs. 2.75 Mb/cM for 2DL; Supplementary figure S2.7a). Data for pericentromeric bins were compiled resulting in a median recombination frequency of 9.1 Mb/cM (Supplementary figure S2.7a). The simulation used real-life set of sizes of the 80 sequence scaffolds. These were randomly picked until the cumulative size had reached the size of the respective chromosome bin. Then, the target gene was positioned randomly inside the bin. Next, the recombination breakpoints were distributed randomly across the chromosome segments (assuming recombination frequency to be evenly distributed along the bin). The number of recombination breakpoints was determined by population size and recombination frequency for the respective bin. Finally, the software tested whether the target gene was flanked by two recombination breakpoints on the same sequence scaffold. The sizes of tested mapping populations ranged from 50–2,000 individuals, increasing the population size in steps of 50. The simulation was repeated 10,000 times for each population size, which provided the probabilities of the gene being flanked on both sides by genetic markers on the same sequence scaffold for different sizes of mapping populations (Supplementary figure S2.7b). All original Perl scripts used for calculations of recombination frequencies and simulations are available upon request.

2.3 Declarations

2.3.1 Data availability

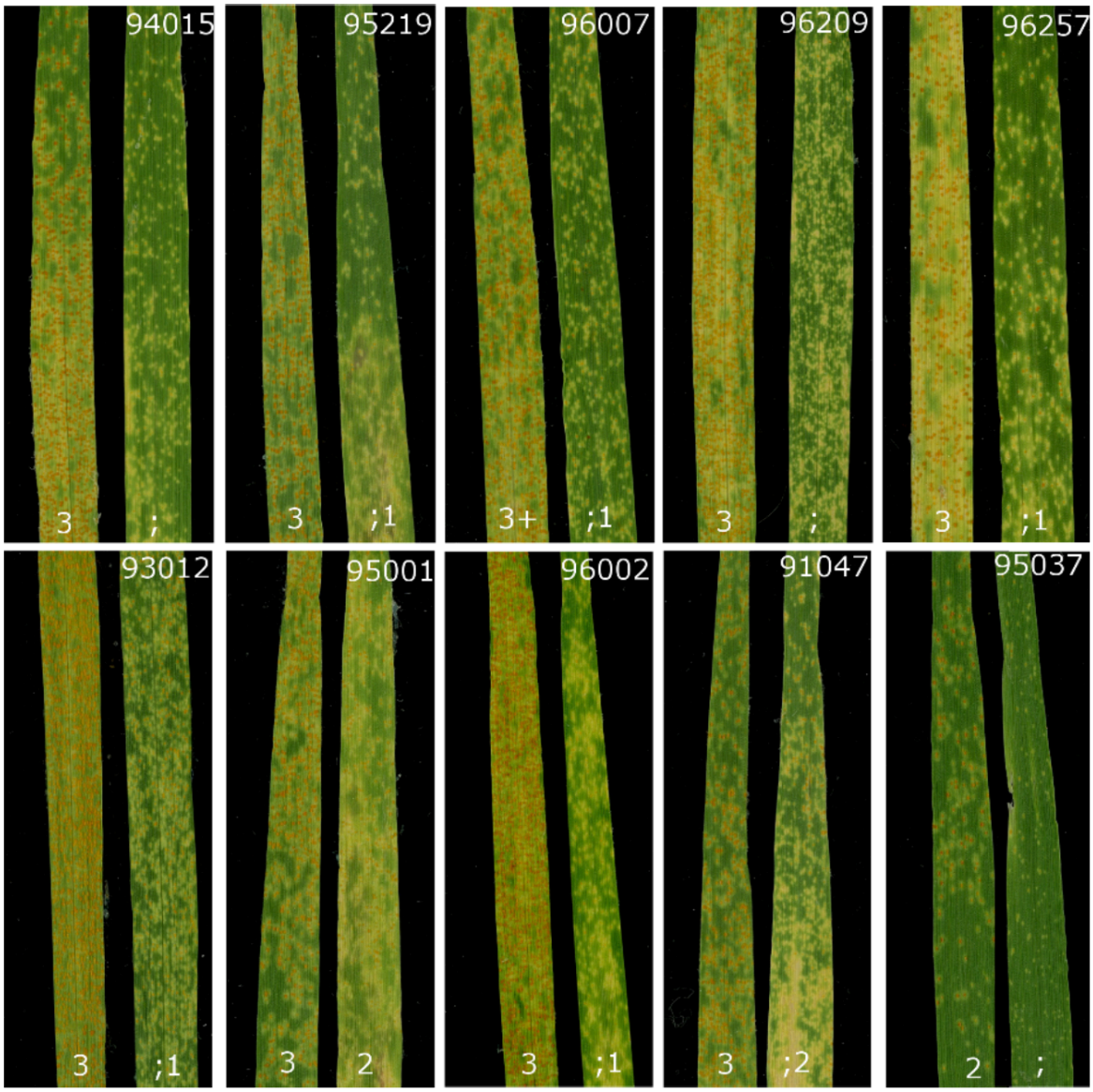
The CH Campala *Lr22a* scaffolds were deposited at DDBJ/ENA/GenBank under the accession MOLT00000000. The version described in this paper is version MOLT01000000.

The *Lr22a* gene sequence was deposited at DDBJ/ENA/GenBank under the accession KY064064. The *NLR1* sequences from Thatcher and Campala have accession numbers KY064065 and KY064066, respectively.

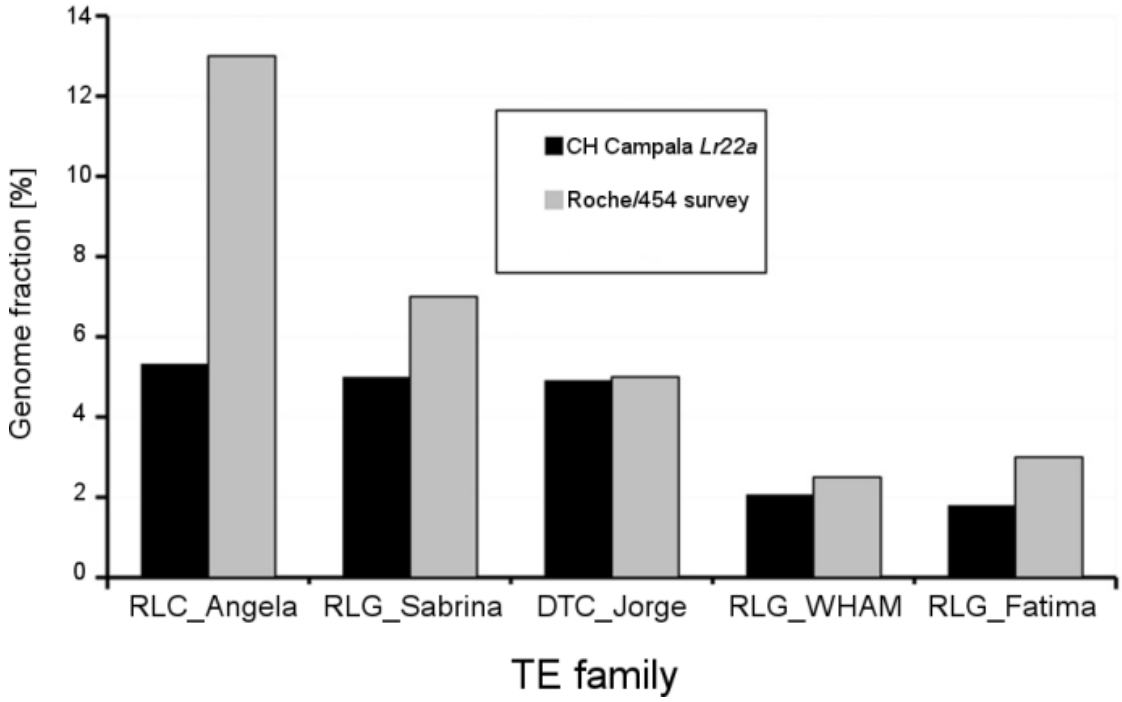
2.3.2 Acknowledgments

We are grateful to the staff at Dovetail Genomics for constructing the CH Campala *Lr22a* scaffolds. We thank M. Karafiátová for supervising chromosome 2D flow sorting and estimation of purity in flow sorted fractions, and Z. Dubská, R. Šperková and J. Weiserová for technical assistance. We also thank B. Senger and L. Luthi for assistance with field experiments and B. Keller for continuous support. This work was financed by an Ambizione fellowship of the Swiss National Science Foundation. J.V., H.Š., and J.D. were supported by the Ministry of Education, Youth and Sports of the Czech Republic (grant award LO1204 from the National Program of Sustainability I).

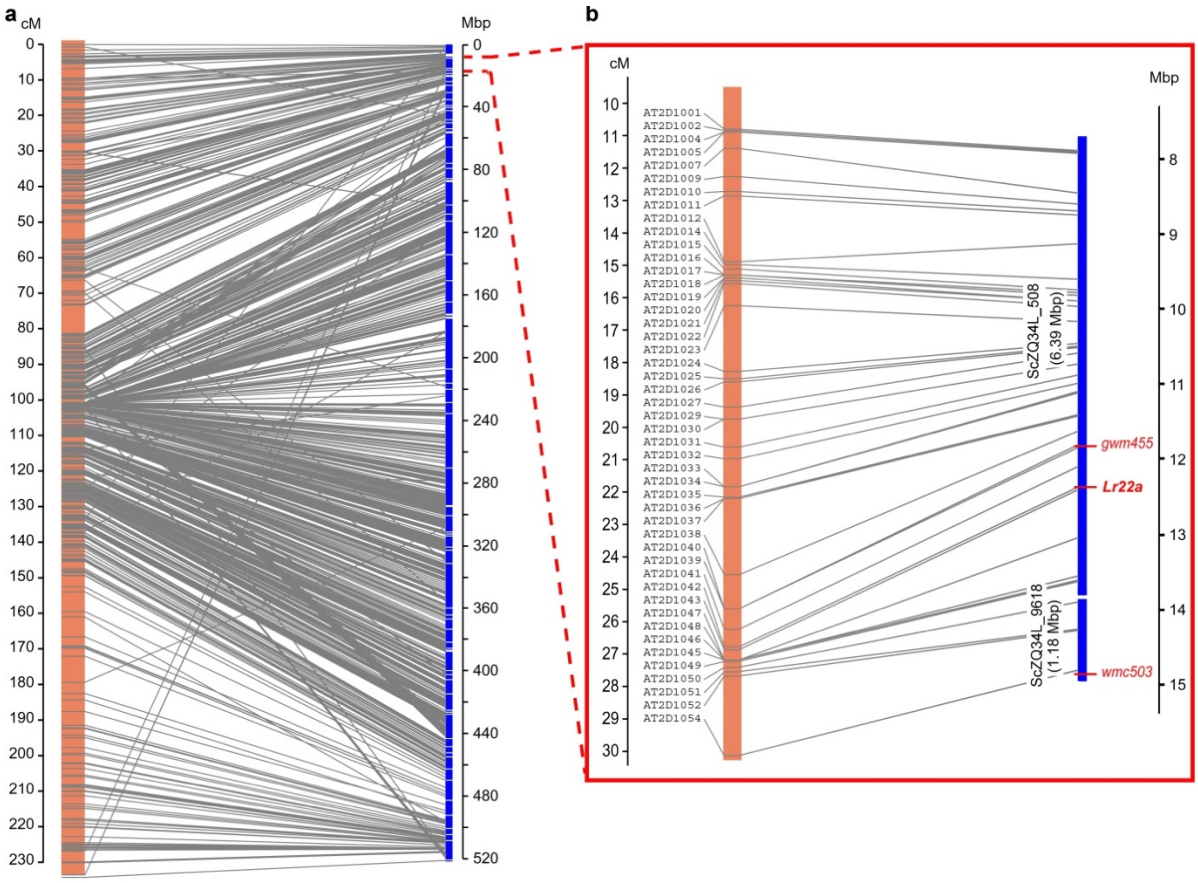
Supplementary figure S2.1. Phenotypic response conferred by the *Lr22a* leaf rust resistance gene against 10 Swiss *P. triticina* isolates. The third leaf of 'Thatcher' (left) and RL6044 (right) is shown 10 days after inoculation. The infection type was scored according to a 0-4 scale (Roelfs, 1984). The isolate number is indicted in the top right corner.



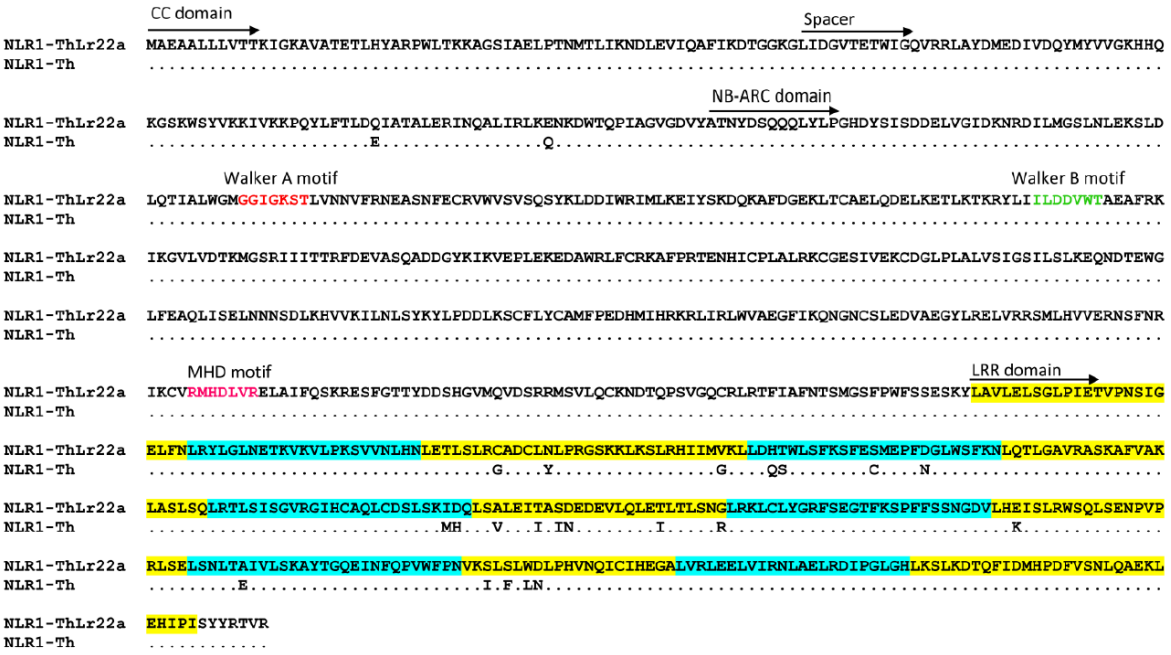
Supplementary figure S2.2. Comparison of transposable element (TE) fraction in the 'CH Campala *Lr22a*' assembly with that of a quantitative survey performed with Roche/454 sequencing (Middleton et al., 2013). For those TE families where data was available, we compared the contributions of annotated TE families. Note that the overall contribution of the high-copy Copia element *RLC_Angela* is much lower in the 'CH Campala *Lr22a*' assembly, indicating that repetitive sequences derived from high-copy TEs are collapsed in the 'CH Campala *Lr22a*' assembly. This may explain why the total length of the 'CH Campala *Lr22a*' assembly was ~160 Mb shorter than the estimated size of chromosome 2D. For this comparison, we annotated 150 Mb (positions 100-250 Mb) of the 'CH Campala *Lr22a*' pseudomolecule ('CH Campala *Lr22a*' scaffolds anchored to genetic *Ae. tauschii* map). The Roche/454 was done on *Ae. tauschii* whole-genome DNA.



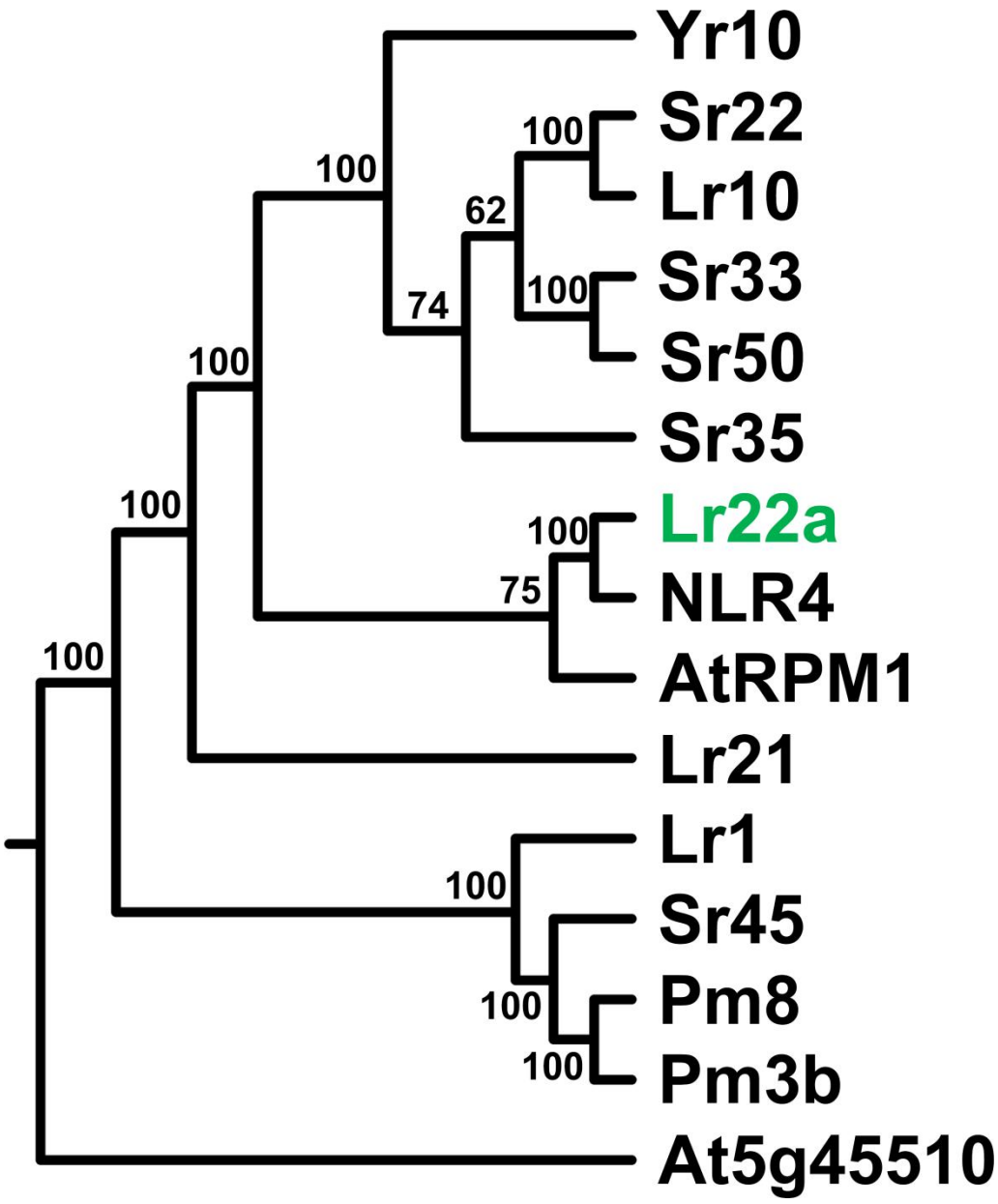
Supplementary figure S2.3. Comparison of the 'CH Campala *Lr22a*' sequence assembly (blue) to the *Ae. tauschii* genetic map (red). (a) Comparison over the entire 2D chromosome and **(b)** the region containing the mapped *Lr22a* markers. The *Lr22a* target interval between markers gwm455 and wmc503 is indicated in red on the 'CH Campala *Lr22a*' assembly.



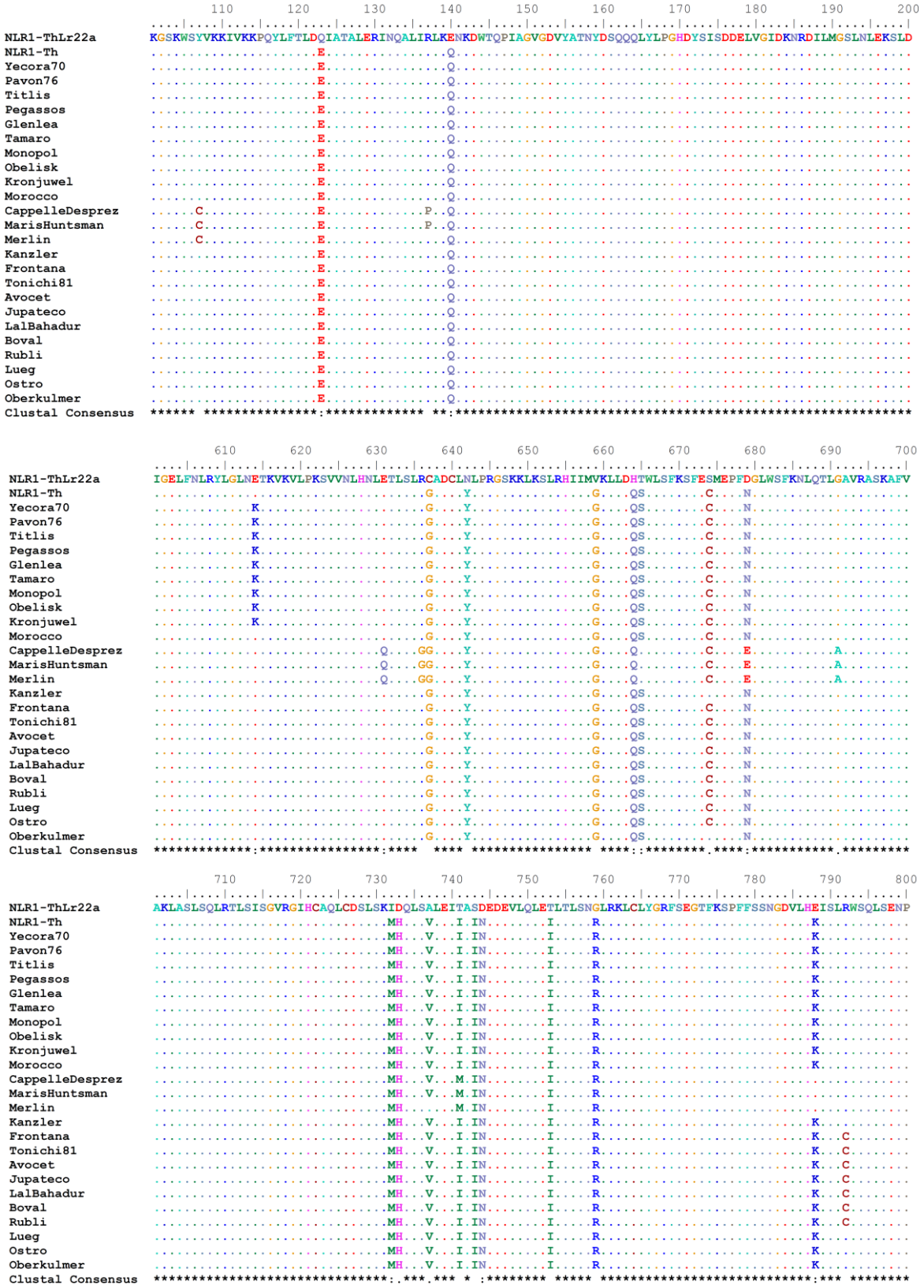
Supplementary figure S2.4. Lr22a protein sequence. The amino acid sequence of Lr22a from RL6044 (NLR1-Th*Lr22a*) is compared to the predicted NLR1 protein version found in the susceptible wheat cultivar 'Thatcher' (NLR1-Th). The *Lr22a* gene sequences in RL6044 and 'CH Campala *Lr22a*' was identical. CC = coiled-coil, NB-ARC = nucleotide-binding, LRR = leucine-rich repeat. The predicted LRR motifs are indicated in yellow and blue, respectively.



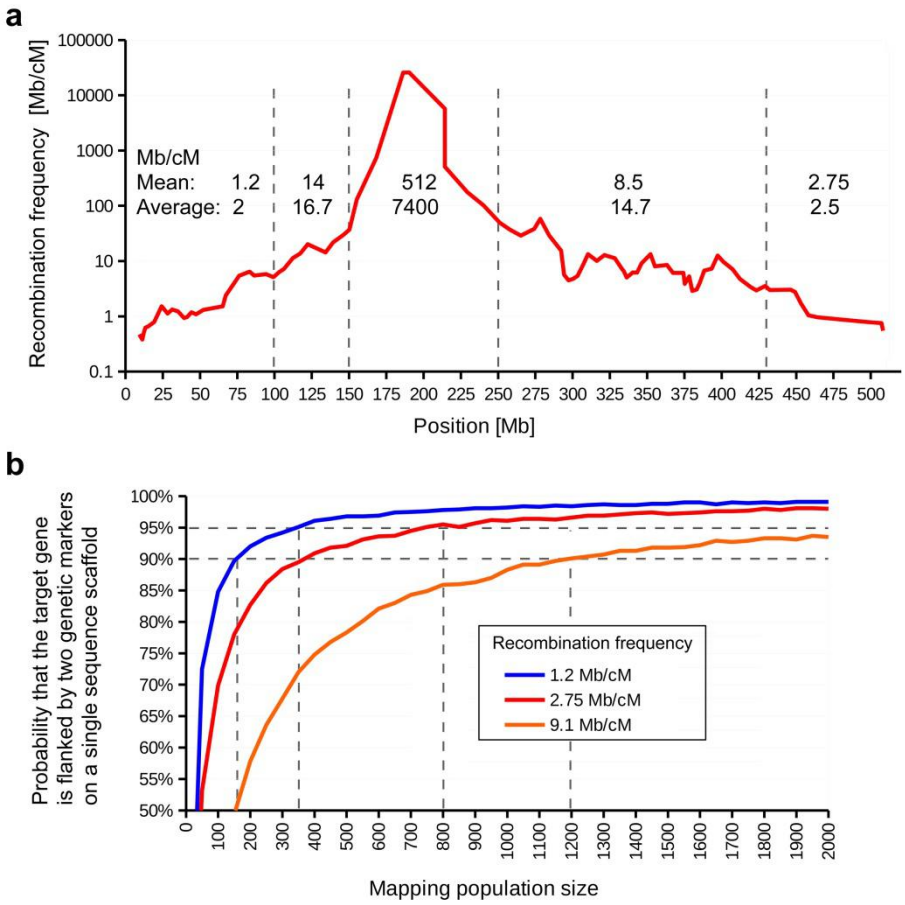
Supplementary figure S2.5. Phylogenetic tree of cloned wheat NLR proteins and RPM1. The LRR domains of the respective proteins were used to construct the tree. Numbers indicate how many times the sequences to the right of the fork occurred in the same group out of 100 trees. The Arabidopsis NLR protein At5g45510 was used to root the tree.



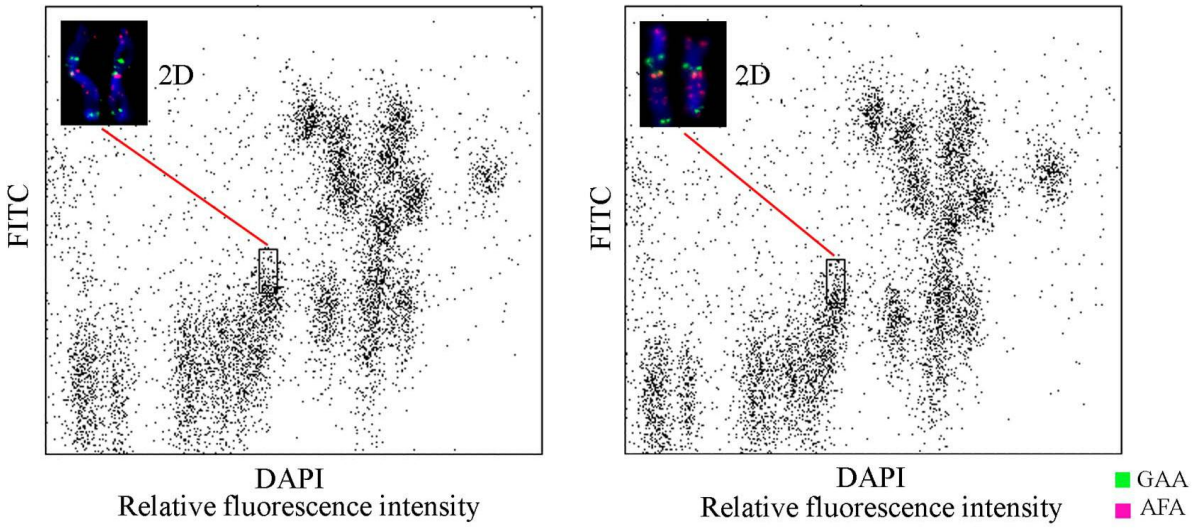
Supplementary figure S2.6. Alignment of Lr22a with NLR1 of 25 wheat cultivars. Shown are the regions that contain unique amino acid (AA) residues in Lr22a in the N-terminal region (AA 123 and 140) and in the LRR region (AA 637-664 and AA 732-756). Ostro and Oberkulmer are spelt wheat accessions.



Supplementary figure S2.7. Simulation of probabilities for a target gene being flanked by two recombination events on a single sequence scaffold. **(a)** Recombination frequencies along chromosome 2D. The x-axis is the position on the 2D pseudomolecule ('CH Campala *Lr22a*' scaffolds anchored to *Ae. tauschii* genetic map) in Mb while the y-axis shows the recombination frequency. Recombination frequencies were based on the *Ae. tauschii* genetic map(Luo et al., 2013) and were calculated in a sliding window for averaging 50 genetic markers at a time. For subsequent simulations, chromosome 2D was separated into two telomeric bins where recombination rates were highest, two pericentromeric bins and one centromeric bin. For each bin, mean and average recombination frequencies are indicated. For chromosome 2D, the telomeric 100 Mb had mean recombination rates of 1.2 Mb/cM for 2DS and 2.75 Mb/cM for 2DL, respectively. Data from chromosome 3B indicate that these two regions may contain well over 60% of the genes(Choulet et al., 2014). **(b)** Simulations to calculate population sizes required for a target gene being flanked by two recombination events on a single sequence scaffold. Simulations are based on the sizes of sequence scaffold used in the 2D pseudomolecule. The dashed lines indicate population sizes necessary to reach 90% or 95% chances of finding a target gene and its closest flanking markers on a single sequence scaffold. Blue = telomeric bin 2DS, red = telomeric bin 2DL, orange = pericentromeric bin (compiled data from both pericentromeric bins).



Supplementary figure S2.8. Bivariate flow karyotypes obtained and the analysis of chromosomes isolated from 'CH Campala' (left) and 'CH Campala *Lr22a*' (right). Sort windows delimiting the populations of chromosome 2D are shown. Insets: Representative images of flow sorted chromosomes 2D which were identified after fluorescence in situ hybridization (FISH) with probes for GAA microsatellites (yellow-green) and *Afa* family repeat (red). Chromosomal DNA was stained by DAPI (blue).



Supplementary table S2.1. Total length, scaffold N50 and N90 of the 'CH Campala *Lr22a*' chromosome 2D assembly and the portion of the assembly that could be anchored to the *Ae. tauschii* genetic map (Luo et al., 2013).

	Total assembly	Anchored to genetic map
Total length	567.2 Mb	521 Mb
N50 length	9.76 Mb (16 scaffolds)	10.11 Mb (13 scaffolds)
N90 length	1.93 Mb (59 scaffolds)	3.87 Mb (44 scaffolds)

Supplementary table S2.2. Molecular characterization of EMS mutants.

Mutant line	Polymorphism	Position of the polymorphism (bp)*	The effect of polymorphism on protein sequence
314	G to A	1,079	amino acid (AA) exchange C360Y
322	G to A	265 and 2,484	AA exchange D89N premature stop codon after AA 827
328	G to A	219	premature stop codon after AA 72
403	G to A	1,463	AA exchange S488N
421	C to T	649	AA exchange L217F

*based on the predicted coding sequence of the *Lr22a* gene.

Supplementary table S2.3. 'CH Campala *Lr22a*' chromosome 2D scaffolds that were anchored to the genetic map of *Ae. tauschii*. The scaffold ID, orientation and scaffold lengths are shown.

scaffold ID	orientation*	length (bp)
ScZQ34L_9148	+	6649149
ScZQ34L_974		4315
ScZQ34L_1032		3690
ScZQ34L_458	-	862050
ScZQ34L_508	+	6391154
ScZQ34L_9618	-	1184130
ScZQ34L_4997	+	1893972
ScZQ34L_252	-	901899
ScZQ34L_187		298942
ScZQ34L_6316	+	2936666
ScZQ34L_7446	-	2185536
ScZQ34L_3386	+	4258924
ScZQ34L_2339	-	4254010
ScZQ34L_3126	+	3461328
ScZQ34L_3179	+	892584
ScZQ34L_1676		1775780
ScZQ34L_2251	+	5519063
ScZQ34L_2831	-	1586397
ScZQ34L_1002	+	3872598
ScZQ34L_7768	-	1908438
ScZQ34L_7665		313614
ScZQ34L_2094	-	8910003
ScZQ34L_669	-	10676775
ScZQ34L_3867	+	2939102
ScZQ34L_6152	-	6592283
ScZQ34L_539		1003721
ScZQ34L_2203		3510
ScZQ34L_8410	+	15510738
ScZQ34L_7295	+	5549600
ScZQ34L_2227	-	4413947
ScZQ34L_10281	-	22151950
ScZQ34L_2202	-	17570666
ScZQ34L_1176	+	14071513
ScZQ34L_1082		7803068
ScZQ34L_182		636186
ScZQ34L_305		1282198
ScZQ34L_4047		185862
ScZQ34L_911	+	33948843

ScZQ34L_5557		9474420
ScZQ34L_8850	+	4002032
ScZQ34L_5198	+	8442372
ScZQ34L_6958	-	7018705
ScZQ34L_2150	-	36420267
ScZQ34L_7561	+	24625224
ScZQ34L_8185		7669
ScZQ34L_2379	-	5763547
ScZQ34L_9756	-	10711663
ScZQ34L_1851	+	1333857
ScZQ34L_5755		798822
ScZQ34L_3277	+	6936978
ScZQ34L_554	-	1778981
ScZQ34L_9544	-	8257848
ScZQ34L_7378	-	29860569
ScZQ34L_5003	-	4429053
ScZQ34L_6345	+	2347547
ScZQ34L_6933	-	6140052
ScZQ34L_1848	-	8788849
ScZQ34L_8682	+	3700333
ScZQ34L_495	-	898356
ScZQ34L_8081		639180
ScZQ34L_9225	+	12304612
ScZQ34L_6768	+	2659125
ScZQ34L_5497	+	2938688
ScZQ34L_2630	+	9367272
ScZQ34L_4900	-	10374543
ScZQ34L_4401		1269428
ScZQ34L_3937	+	1005940
ScZQ34L_722	-	15867027
ScZQ34L_3416	-	5041999
ScZQ34L_617	+	3483752
ScZQ34L_3919	+	2431424
ScZQ34L_7697	+	7698800
ScZQ34L_6030	+	6941149
ScZQ34L_3329	-	13398268
ScZQ34L_2334	+	9758700
ScZQ34L_7673	-	8326716
ScZQ34L_6986	-	4044051
ScZQ34L_2849	-	3880022
ScZQ34L_1222	-	12431682
ScZQ34L_1581		977539

* scaffolds without a + or – could not be oriented because they contained only one SNP marker or several co-segregating SNP markers.

Supplementary table S2.4. List of primers used in this study along with Corresponding SNP marker and scaffold of *Ae. tauschii* (Jia et al., 2013; Luo et al., 2013) from which the respective marker was developed (Methods).

Primer name	Sequence	Use in this work	Amplicon size (bp)	Polymorphism	Annealing temp. [°C]	SNP marker*	*Ae. tauschii* scaffold*
SWSNP4_F	GTGCGACGCCGACCTGATG	Map *Lr22a*	313	G to A at 166 bp	55		
SWSNP4_R	CTGGCTGACGATGATCCG						
SWInDel4_F	GAATTGATGGGCTCGACTAC	Map *Lr22a*	196	6 bp deletion in Campala	55	AT2D1042	Atau_2D_4341.2
SWInDel4_R	CGCAGCACATCTGGTGGG						
SWSNP5_F	GACTGATCAGACTATGG	Map *Lr22a*	178	A to T at 130 bp	55	AT2D1042	Atau_2D_3080.2
SWSNP5_R	CCAATTCACGTACAAGATC						
SWSNP6_F	CATCATGGCCGACCACGCC	Map *Lr22a*	187	C to T at 64 bp	60		
SWSNP6_R	CTCCGGTGCACCGTGGAG						
SWInDel7_F	GACCTAGGGATACGCGCATG	Map *Lr22a*	550		55		
SWInDel7_R	GGTTCAGTATACGTACGAG						
LRR1-F3	CAT AGC ATC ATT CGC GAG AC	Amplification of *Lr22a*	3227		63		
LRR1-R4	CAA GCA TAC ACT GAA CAG C						
LRRSEQ-R7	GAT TGA GTA ATC ATG TCC AG	Sequencing of *Lr22a*					
LRR1SEQ-F8	GCT ACT GCC CTT GAG AG	Sequencing of *Lr22a*					
LRRSEQ-F9	CTA GAT GAT ATT TGG AG	Sequencing of *Lr22a*					
LRRSEQ-F10	GGA GGA TGC ATG GCG TC	Sequencing of *Lr22a*					
LRRSEQ-F11	GAT GTT GCT GAA GGT TAC	Sequencing of *Lr22a*					
LRRSEQ-F12	GGA TTA CCT ATT GAG ACT	Sequencing of *Lr22a*					
LRRSEQ-F13	GGT GCA GTT CGA GCT AG	Sequencing of *Lr22a*					
LRRSEQ-F14	GCA ACG GAG ATG TCC TTC AC	Sequencing of *Lr22a*					

Chapter 3

Chromosome-scale comparative sequence analysis unravels molecular mechanisms of genome dynamics between two wheat cultivars

Anupriya Kaur Thind[+], Thomas Wicker[+], Thomas Müller, Patrick M. Ackermann, Burkhard Steuernagel, Brande B.H. Wulff, Manuel Spannagl, Sven O. Twardziok, Marius Felder, Thomas Lux, Klaus F.X. Mayer, International Wheat Genome Sequencing Consortium, Beat Keller, Simon G. Krattinger

[+] Authors contributed equally to this work

(2018)

Genome Biology 19:104

doi: 10.1186/s13059-018-1477-2

Abstract

Recent improvements in DNA sequencing and genome scaffolding have paved the way to generate high-quality de novo assemblies of pseudomolecules representing complete chromosomes of wheat and its wild relatives. These assemblies form the basis to compare the dynamics of wheat genomes on a megabase-scale. Here, we provide a comparative sequence analysis of the 700 megabase chromosome 2D between two bread wheat genotypes – the old landrace Chinese Spring and the elite Swiss spring wheat line 'CH Campala *Lr22a'*. Both chromosomes were assembled into megabase-sized scaffolds. There is a high degree of sequence conservation between the two chromosomes. Analysis of large structural variations reveals four large indels of more than 100 kb. Based on the molecular signatures at the breakpoints, unequal crossing over and double-strand break repair were identified as the molecular mechanisms that caused these indels. Three of the large indels affect copy number of NLRs, a gene family involved in plant immunity. Analysis of SNP density reveals four haploblocks of 4 Mb, 8 Mb, 9 Mb and 48 Mb with a 35-fold increased SNP density compared to the rest of the chromosome. Gene content across the two chromosomes was highly conserved. Ninety-nine percent of the genic sequences were present in both genotypes and the fraction of unique genes ranged from 0.4 to 0.7%. This comparative analysis of two high-quality chromosome assemblies enabled a comprehensive assessment of large structural variations and gene content. The insight obtained from this analysis will form the basis of future wheat pan-genome studies.

Keywords: genome diversity, structural variation, high-quality assembly, wheat

3.1 Introduction

Bread wheat (*Triticum aestivum*) was the most widely grown cereal crop in 2016. It serves as a staple food for over 30% of the world's population and provides ~20% of the globally consumed calories (FAO, 2017). Wheat is a young allopolyploid species with a genome size of 15.4-15.8 Gb, of which more than 85% is made up of highly repetitive sequences (Wicker et al., 2018). The allopolyploid genome arose through two recent, natural polyploidization events that involved three diploid grass species. The first hybridization event occurred 0.58 to 0.82 million years ago (Jordan et al., 2015) between the A-genome donor wild einkorn (*T. urartu*) and a yet unidentified B-genome donor that was a close relative of *Aegilops speltoides*. This hybridization created wild tetraploid emmer wheat (*Triticum turgidum* ssp. *dicoccoides*; AABB genome) (Avni et al., 2017). A second natural hybridization between domesticated emmer and wild goatgrass (*Ae. tauschii;* DD genome) resulted in the formation of hexaploid bread wheat (AABBDD genome) around 10,000 years ago (Salamini et al., 2002). The domestication of tetraploid emmer and the limited number of hybridization events with *Ae. tauschii* represent bottlenecks that resulted in a significant reduction of genetic diversity within the bread wheat gene pool. Natural gene flow between bread wheat and its wild and domesticated relatives as well as artificial hybridizations with diverse grass species partially compensated for this loss in diversity (Akhunov et al., 2010; Jordan et al., 2015).

The size, repeat content and polyploidy of the bread wheat genome have represented major challenges for the generation of a high-quality reference assembly. The first 'early' whole genome assemblies of hexaploid wheat and its diploid wild relatives were based on short-read sequencing approaches. These assemblies provided an insight into the gene space of wheat, but they were highly fragmented and incomplete (Brenchley et al., 2012; International Wheat Genome Sequencing Consortium, 2014; Jia et al., 2013; Ling et al., 2013). The first notable high-quality sequence assembly of wheat was produced from the 1-gigabase chromosome 3B of

the hexaploid wheat landrace Chinese Spring. For this, 8,452 ordered bacterial artificial chromosomes (BACs) were sequenced and assembled, which resulted in a highly contiguous assembly (N50 = 892 kb) (Choulet et al., 2014; Paux et al., 2008). More recent whole-genome shotgun assemblies had improved contiguousness compared to the 'early' assemblies (N50 = 25 – 232 kb) (Chapman et al., 2015; Clavijo et al., 2017; Zimin et al., 2017), but they still did not allow to compare the structure of wheat chromosomes on a megabase-scale.

Several recent technological and computational improvements however provided a basis to generate *de novo* assemblies of complex plant genomes with massively improved scaffold lengths and completeness. These advancements included (*i*) the integration of whole-genome shotgun libraries of various insert-sizes (Hirsch et al., 2016) or the use of long-read sequencing technologies such as single-molecule real-time sequencing (SMRT) (Jiao et al., 2017) or nanopore sequencing (Schmidt et al., 2017), (*ii*) the improvement of scaffolding by using chromosome conformation capture technologies (Jarvis et al., 2017; Lieberman-Aiden et al., 2009; Mascher et al., 2017; Putnam et al., 2016; van Berkum et al., 2010) or optical maps (Moll et al., 2017) and (*iii*) the improvement of assembly algorithms (Avni et al., 2017). With the use of some of these novel approaches, a near complete reference assembly of Chinese Spring (IWGSC RefSeq v1.0) with a scaffold N50 of 22.8 Mb was recently generated (International Wheat Genome Sequencing Consortium, 2018). Chinese Spring is an old landrace that was selected for sequencing because it was used in a number of cytogenetic studies, which has resulted in the generation of many important genetic resources from this wheat line, including chromosome deletion lines (Endo & Gill, 1996) and aneuploid lines (Sears & Sears, 1978).

The completion of the IWGSC RefSeq v1.0 assembly lays the foundation to study the genetic diversity within and between different wheat species and cultivars. The understanding of this genetic variation will provide an insight into wheat genome dynamics and its impact

on agronomically important traits. The continuum of genetic variation ranges from single nucleotide polymorphism (SNPs) to megabase-sized rearrangements that can affect the structure of entire chromosomes (Alkan et al.; 2011). Due to the absence of high-quality wheat genome assemblies, previous comparative analyses were limited in the size of structural rearrangements that could be assessed and typically, structural variants of a few base pairs up to several kb were analysed (Liu et al.; Montenegro et al., 2017). Consequently, a comprehensive assessment of the extent of large structural rearrangements and their underlying molecular mechanisms is still lacking.

Here, we report on a chromosome-wide comparative analysis of the ~700 Mb chromosome 2D between the two hexaploid wheat lines Chinese Spring and 'CH Campala *Lr22a*'. 'CH Campala *Lr22a*' is a backcross line that was generated to introgress *Lr22a,* a gene that provides resistance against the fungal leaf rust disease, into the genetic background of the elite Swiss spring wheat cultivar 'CH Campala' (Moullet et al.; 2014). We previously generated a high-quality *de novo* assembly from isolated chromosome 2D of 'CH Campala *Lr22a*' by using short-read sequencing in combination with Chicago long-range scaffolding (Thind et al., 2017). The resulting assembly had a scaffold N50 of 9.76 Mb. Here, we compared this high-quality assembly to chromosome 2D of the Chinese Spring IWGSC RefSeq v1.0 assembly. In particular, the focus of our study was on the identification and quantification of large structural variations (SVs). The comparative analysis of the 2D chromosome showed a high degree of collinearity along most of the chromosome, but also revealed SVs such as InDels and copy number variation (CNV). In addition, we found haploblocks with greatly increased SNP densities. We analysed these SVs and gene presence/absence polymorphisms in detail and manually validated them to distinguish true SVs from artefacts that were due to mis-assembly or annotation problems.

3.2 Results

3.2.1 Two-way comparison of Chinese Spring and 'CH Campala *Lr22a*' allows identification of large structural variations

Previously, 10,344 sequence scaffolds were produced from isolated chromosome 2D of 'CH Campala *Lr22a*' by using Chicago long-range linkage (Putnam et al., 2016; Thind et al., 2017). To construct a 'CH Campala *Lr22a*' pseudomolecule, we anchored these scaffolds to the IWGSC RefSeq v1.0 chromosome 2D using BLASTN (see methods). In the resulting 'CH Campala *Lr22a*' pseudomolecule, 7,617 scaffolds were anchored, of which 7,314 were smaller than 5 kb and 90 scaffolds were larger than 1 Mb in size. The pseudomolecule had a scaffold N50 of 8.78 Mb (N90 of 1.89 Mb) and represented 98.92% of the total length of the initial assembly. The 'CH Campala *Lr22a*' pseudomolecule has a total length of 563 Mb whereas the IWGSC RefSeq v1.0 2D pseudomolecule is 651 Mb in length. It was previously found that repetitive sequences were collapsed and less complete in the Chicago assembly, which explains the smaller size of the 'CH Campala *Lr22a*' pseudomolecule compared to the IWGSC RefSeq v1.0 pseudomolecule (Thind et al., 2017). In total, 6,018 high confidence (HC) genes were annotated in Chinese Spring (International Wheat Genome Sequencing Consortium, 2018) and 5,883 HC genes in 'CH Campala *Lr22a*' (see methods). Of the 5,883 'CH Campala *Lr22a*' HC genes, 45 genes were located on short scaffolds that contained no other gene. Gene annotation and collinearity will be discussed in detail in a following paragraph.

To identify large InDels, we compared the Chinese Spring and 'CH Campala *Lr22a*' pseudomolecules in windows of 10 Mb and performed dot plots. Here, we focused only on InDels larger than 100 kb because such SVs could not be identified with previous whole-genome assemblies. In total, we found 26 putative InDels which were manually validated by

evaluating the upstream and downstream sequences for the presence of 'Ns' at the breakpoints. If 'Ns' were found exactly at the breakpoints on both sides of an InDel, we considered it a false positive that was most likely due to the incorrect placement of a scaffolds in either of the pseudomolecules. Based on this criterion, we discarded 22 of the 26 candidate InDels. Three of the remaining four InDels showed good sequence quality and had clear breakpoints at both ends with no 'Ns'. These true InDels were 285 kb, 494 kb and 765 kb in size. An additional 677 kb InDel had a clear break only at one end and 'Ns' on the other end. Interestingly, three of the four large InDels showed CNV for nucleotide binding site – leucine-rich repeat (NLR) genes.

Various molecular mechanisms have been described that lead to SVs. For example, unequal crossing over can occur in regions with extensive sequence similarity. On the other hand, non-homologous end-joining (NHEJ) is associated with DNA repair in regions with no or low sequence similarity. Other causes of SVs include double-strand break (DSB) repair via single-strand annealing or synthesis-dependent strand annealing mechanisms, transposable element (TEs)-mediated mechanisms and replication-error mechanisms (Munoz-Amatriain et al., 2013; Robberecht, 2013; Wicker et al., 2010; T. Wicker et al., 2016). These mechanisms have been well studied in humans, but in plants our understanding of the molecular causes of SVs is limited (Munoz-Amatriain et al., 2013). To decipher the mechanistic bases of the observed SVs, the sequence of the SV as well as their flanking regions were analyzed to identify signature sequence motifs that could point to the underlying molecular mechanism (e.g. DNA repair, recombination or replication associated mechanisms).

3.2.1 Unequal crossing over is the likely cause of a 285 kb deletion in Chinese Spring

Sequence comparison revealed an InDel of 285 kb on the short chromosome arm (Fig. 3.1a). We extracted and checked the sequences 5 kb upstream and downstream of the

breakpoints for the presence of TEs or genes (or any kind of repeated sequence) that could have served as a template for unequal crossing over. Unequal crossing over occurs frequently at repeated sequences that are in the same orientation, leading to duplications or deletions of the region between the two repeats (Cai & Xu, 2007). Indeed, the breakpoints of the InDel contained two NLR genes that shared 96-98% nucleotide identity in 'CH Campala *Lr22a*'. In contrast, Chinese Spring only carried a single NLR copy (Fig. 3.1). Thus, it is possible that an unequal crossing over between the two genes occurred in an ancestor of Chinese Spring, leading to the loss of the 285 kb segment between the two NLRs.

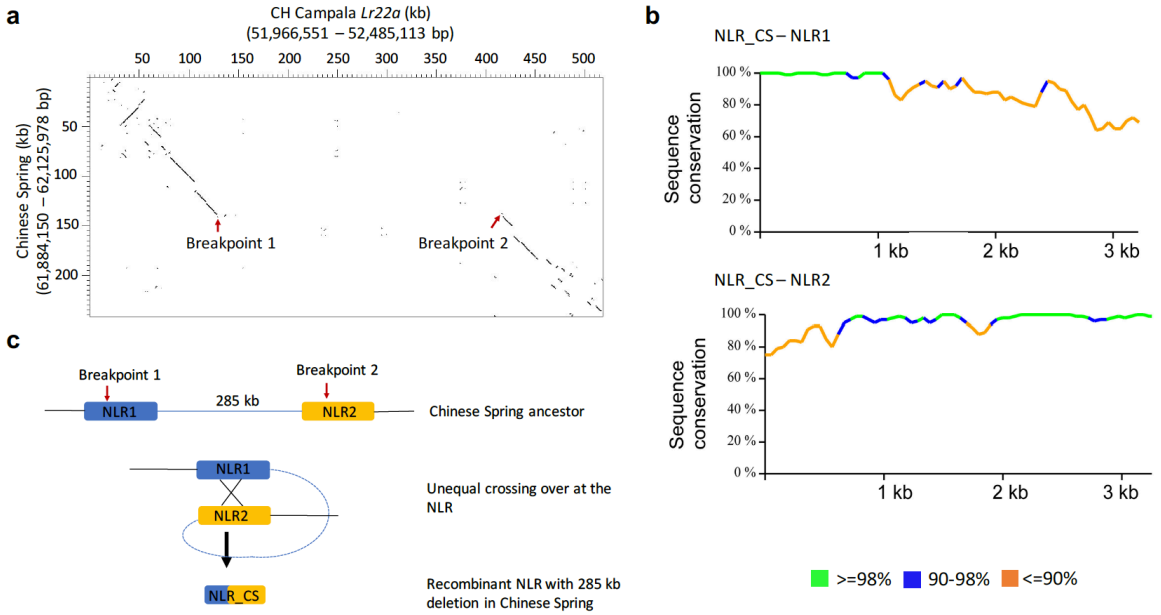


Fig. 3.1 Unequal crossing over resulted in a 285 kb deletion in Chinese Spring. a Dot plot of a 525 kb segment from 'CH Campala *Lr22a*' against the corresponding 280 kb segment from Chinese Spring. The breakpoints of the 285 kb deletion are indicated by red arrows. The numbers in brackets refer to the positions of the selected region on the respective pseudomolecule. **b** Pairwise alignment of the Chinese Spring NLR with the two 'CH Campala *Lr22a*' NLRs shows putative recombination breakpoints that led to the formation of the Chinese Spring NLR. **c** Proposed model for molecular events that led to a 285 kb deletion in Chinese Spring. An unequal crossing over event involving two NLR genes (shown in blue and orange) led to the formation of the recombinant NLR in Chinese Spring which shares sequence homology with NLR1 (blue) and NLR2 (yellow) and a deletion of the intervening 285 kb sequence.

In order to test this hypothesis, we further analysed the NLRs that were present at the breakpoint of 'CH Campala *Lr22a*' and Chinese Spring. Interestingly, the 5' region of the

Chinese Spring gene showed greater sequence similarity to NLR1 of 'CH Campala *Lr22a*', whereas the 3' region was more similar to NLR2 (Fig. 3.1b). This suggests that these NLRs (NLR1 and NLR2) were indeed the template for an unequal crossing over in an ancestor of Chinese Spring (Fig. 3.1c). The corresponding 285 kb segment in 'CH Campala *Lr22a*' only contained repetitive sequences and did not carry any genes.

3.2.2 Double-strand break repair likely mediated a large 494 kb deletion

The second SV was located on a 'CH Campala *Lr22a*' scaffold of 6.6 Mb in size (Fig. 3.2a). We could precisely identify the breakpoints based on the sequence alignment of the two wheat lines. Unlike the case described above, the upstream and downstream sequences contained no obvious sequence template or a typical TE insertion or excision pattern (Wicker et al., 2010) that could have led to a large deletion by unequal crossing over. However, the breakpoints of the InDel contained typical signatures of DSB repair. In 'CH Campala *Lr22a*' the nucleotide triplet 'CGA' was repeated at both ends of the breakpoint whereas Chinese Spring had only one copy of the 'CGA' triplet (Fig. 3.2b). The proposed model for this 494 kb deletion is that it was caused through a DSB that was repaired by the single-strand annealing pathway (Fig. 3.2c). After the DSB that could have occurred anywhere on the 494 kb segment in Chinese Spring, 3' overhangs were produced by exonucleases. Various studies in yeast have shown that these overhangs can be many kilobases in size (Fishman-Lobell Jacqueline, 1992; Storici, Snipe, Chan, Gordenin, & Resnick, 2006; Yang, Sterling, Storici, Resnick, & Gordenin, 2008) and due to high conservation of DSB repair pathways (Shevelev & Hubscher, 2002), it is expected that plants would have a similar DSB repair mechanism. In the case described here, we propose that exonucleases produced overhangs of 200-250 kb, which were then repaired by non-conservative homologous recombination repair (HRR). For this, the generated 3' overhangs annealed in a place of complementary micro-homology, which are typically a few bp in size ('CGA' triplet in this case)

(Pfeiffer et al., 2000). After annealing of the matching motifs, second strand synthesis took place and the overhangs were removed, leading to the observed deletion of the 494 kb sequence in Chinese Spring (Fig. 3.2c). This 494 kb segment in 'CH Campala *Lr22a*' contained eight genes coding for an NLR, a serine/threonine protein kinase, a zinc finger-containing protein, a transferase, two cytochrome P450s and two proteins of unknown function. BLAST analysis of these eight genes against the IWGSC RefSeq v1.0 pseudomolecules revealed that the homoeologous segments on the A and B genomes were retained. In other words, the deletion of these eight genes might not have led to a deleterious effect because the homoeologous gene copies on the other two sub-genomes compensate for the D-genome deletion. It has been reported that polyploid species show a higher plasticity compared to diploid species and that they are able to buffer large insertions and deletions on one particular sub-genome (Leitch & Leitch, 2008).

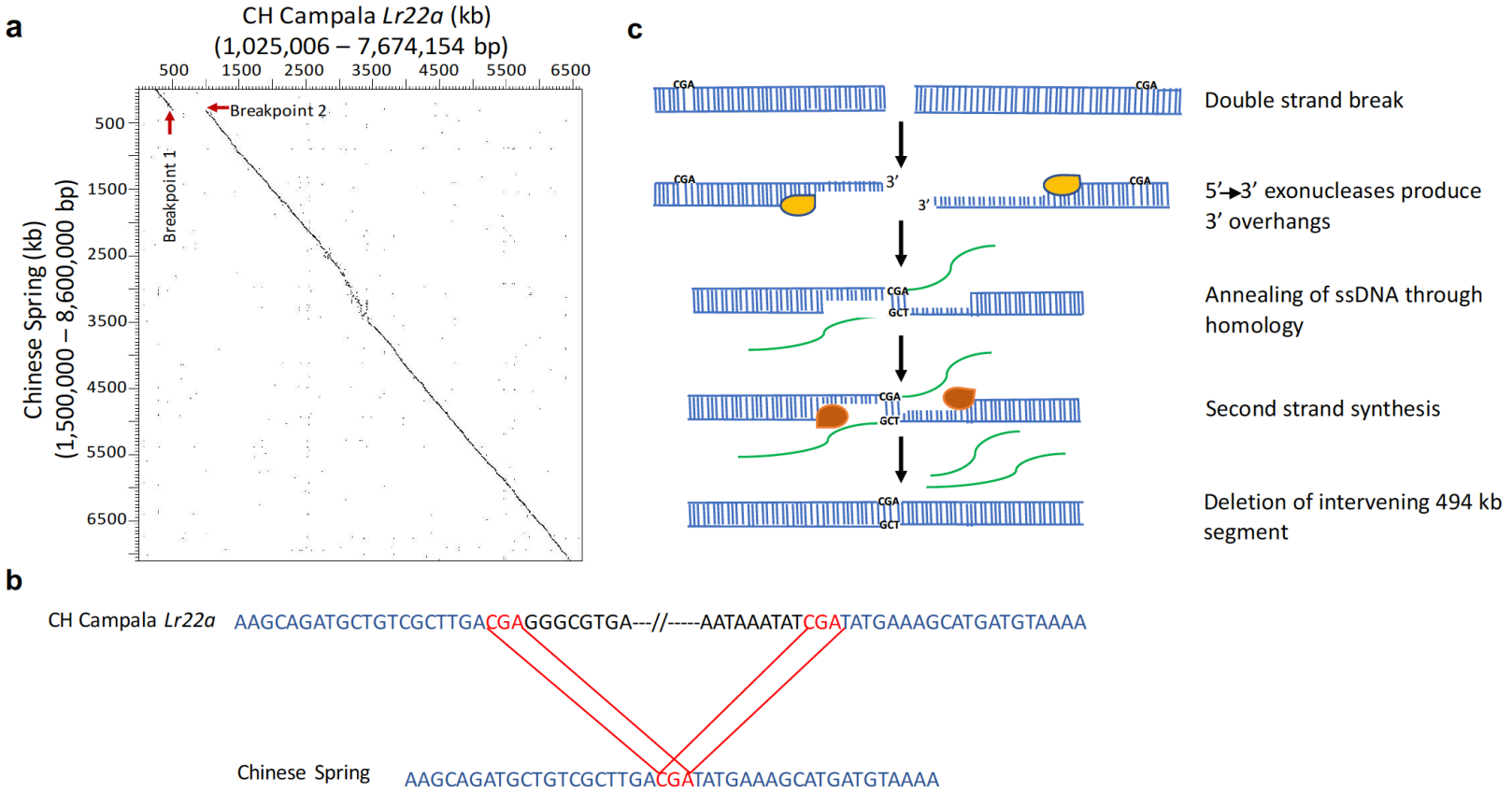


Fig. 3.2 Double-strand break repair is responsible for the deletion of a 494 kb segment in Chinese Spring. a Dot plot of a 6.6 Mb scaffold of 'CH Campala *Lr22a*' against the corresponding segment from Chinese Spring. The breakpoints are indicated by red arrows. The numbers in brackets refer to the positions of the selected region on the respective pseudomolecule. **b** Presence of DSB signatures ('CGA' triplet, red) with two copies in 'CH Campala *Lr22a*' and one in Chinese Spring. The conserved sequence is shown in blue and the 494 kb sequence that is deleted in Chinese Spring but present in 'CH Campala *Lr22a*' is indicated in black. **c** The proposed model for the deletion of the 494 kb segment in Chinese Spring through DSB repair by the single-strand annealing pathway, where the yellow enzyme is the exonuclease, green strands are the overhangs and the orange color represents the replication complex.

3.2.3 Large diverse haploblocks indicate recurrent gene flow from distant relatives

Comparison of SNP density across the chromosome revealed four large regions (haploblocks *a*, *b*, c and *d*) with increased SNP density compared to the rest of the chromosome (Fig. 3.3a). Two of the regions were located on the short arm of the chromosome whereas the largest diverse haploblock of ~48 Mb and a shorter fourth haploblock were located towards the telomeric end of the long chromosome arm. While the SNP density along most of the chromosome was in the range ~27 SNPs/Mb (Fig. 3.3a) the four diverse haploblocks had SNP densities of 2,500 – 4,500 SNPs/Mb. The actual number of polymorphisms might be even higher because SNP calling might not have been possible in many parts of the haploblocks because of the high sequence divergence.

The first haploblock (haploblock *a*) at the distal end of the short chromosome arm contains the *Lr22a* leaf rust resistance gene that was introduced into hexaploid wheat through an artificial hybridization between a tetraploid wheat line and an *Ae. tauschii* accession (Dyck & Kerber, 1970). There are two genetically distant lineages of *Ae. tauschii*. The D-genome of hexaploid wheat was most likely contributed by an *Ae. tauschii* population belonging to lineage 2 (Wang et al., 2013), whereas the donor of *Lr22a* (*Ae. tauschii* accession RL 5271) belongs to the genetically diverse lineage 1 (Arora et al., 2017). The size of the *Lr22a* introgression was subsequently reduced through several rounds of backcrossing with hexaploid wheat and the remaining *Lr22a*-containing segment was bred into elite wheat lines including 'CH Campala *Lr22a*' to increase resistance against the fungal leaf rust disease (Moullet et al., 2014). Based on the SNP density, we were able to estimate the size of the remaining, introgressed *Ae. tauschii* segment to ~8 Mb. The original donor of the other three haploblocks (haploblocks *b*, *c* and *d*) could not be traced back and they might be the result of natural gene flow or artificial hybridization. Mapping of independently generated short-read sequences from 'CH Campala', the recurrent parent that was used to produce the near isogenic line 'CH Campala *Lr22a*', showed

that the same haploblocks were also present in 'CH Campala' (Fig. 3.3a), indicating that these segments were not co-introduced along with the *Lr22a* segment from RL 5271. Haploblock *b* comprised the 285 kb deletion described above (Fig. 3.1). In particular, the presence of the large continuous haploblock *c* on the long chromosome arm was intriguing. Dot plots allowed us to identify the exact breakpoints of the haploblock (Fig. 3.3b). While there was high sequence homology in both flanking regions, sequence identity in the intergenic regions broke down inside the haploblock (Fig. 3.3b). In contrast, dot plots with haploblocks *a, b* and *d* revealed a good level of collinearity between Chinese Spring and 'CH Campala *Lr22a*' in intergenic regions despite the increased SNP density (Supplementary figure S3.1), indicating that haploblock *c* is the most diverse. Comparison to the recently generated high-quality genome assembly of *Ae. tauschii* accession AL8/78 (Luo et al., 2017), an accession that is closely related to the wheat D-genome and that belongs to lineage 2, suggests that haploblock *c* represents an interstitial introgression into 'CH Campala *Lr22a*' (Supplementary figure S3.2). In Chinese Spring, 723 genes were located in this haploblock, whereas 'CH Campala *Lr22a*' contained 678 genes in this region (Supplementary table S3.1). The genic sequences in the haploblock region showed a nucleotide sequence identity of 78-100% compared to 99-100% for the genes outside the haploblock. We also observed three inversions of ~1.48 Mb, ~422 kb and ~418 kb in the haploblock *c* where the gene order was reversed. To track the possible origin of this introgression, we developed an introgression-specific PCR probe based on the sequence of the left breakpoint in 'CH Campala *Lr22a*'. The marker amplified in several wheat cultivars that were developed by the International Wheat and Maize Improvement Center (CIMMYT) (Fig. 3.3c). Among them is Inia-66, which is in the pedigree of 'CH Campala *Lr22a*' (http://www.wheatpedigree.net/sort/show/118822). These results indicate that the particular segment in 'CH Campala *Lr22a*' might have been introgressed via a CIMMYT cultivar.

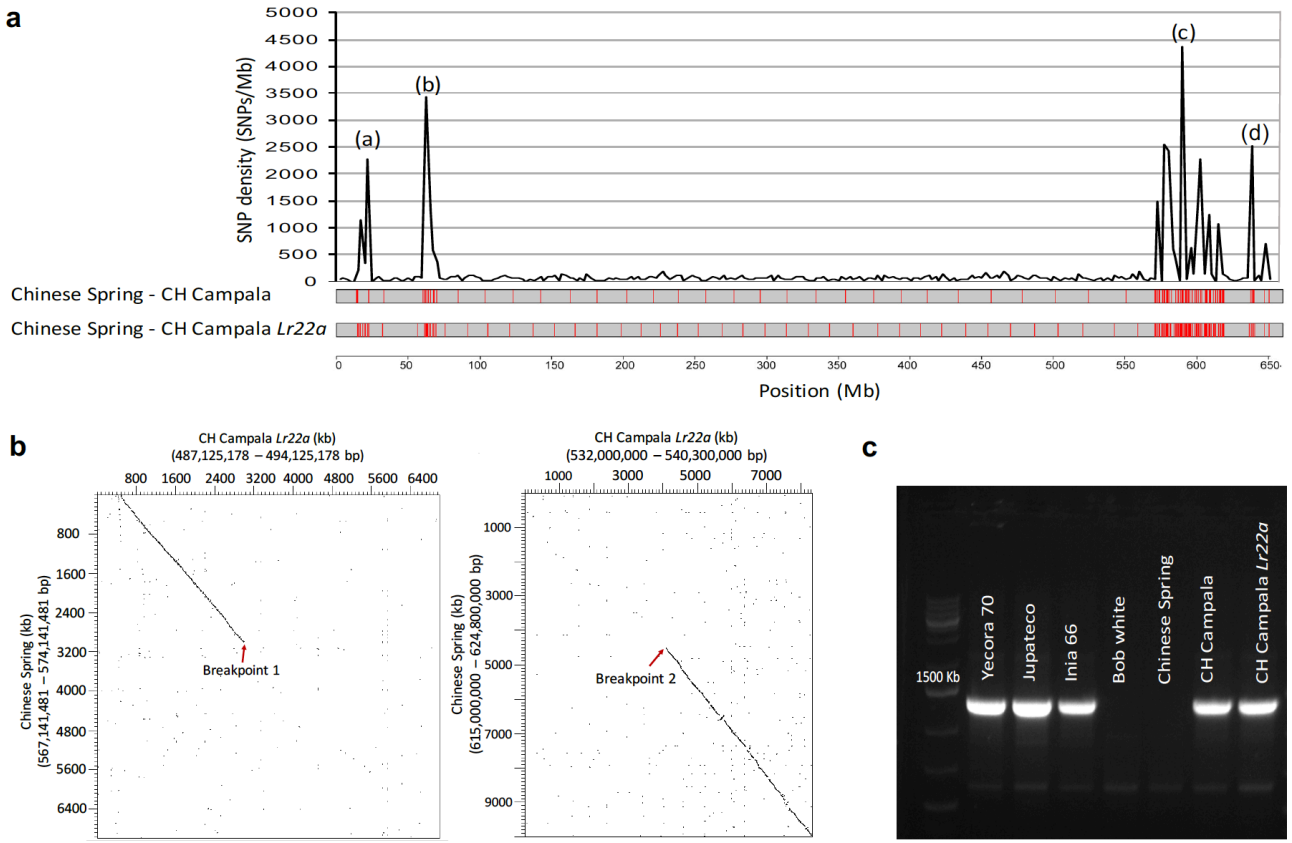


Fig. 3.3 Identification of three diverse haploblocks with increased SNP density. a Single nucleotide polymorphism (SNP) density between Chinese Spring and 'CH Campala *Lr22a*' in a sliding windows of 2.5 Mb. The numbers refer to the position in Mb along the chromosome 2D of the Chinese Spring. The three diverse haploblocks are indicated with letters (a), (b) and (c). **b** Dot plot of Chinese Spring and 'CH Campala *Lr22a*' showing the left and right breakpoints of the large haploblock *c*. The sequence adjacent to the haploblock shows a high degree of sequence conservation in intergenic regions whereas the sequence similarity was very low in the haploblock region. **c** PCR amplification using an introgression specific primer designed on the left breakpoint of the 'CH Campala *Lr22a*' introgression. Jupateco, Yecora 70 and Inia 66 are CIMMYT wheat cultivars. Inia 66 is in the pedigree of 'CH Campala *Lr22a*'.

3.2.4 Presence of unique genes and gene synteny

A total of 6,018 high confidence (HC) genes were annotated on chromosome 2D of the Chinese Spring reference sequence (IWGSC v1.0; (International Wheat Genome Sequencing Consortium, 2018)) and 5,883 HC genes were annotated on chromosome 2D of 'CH Campala *Lr22a*'. A BLASTN analysis of the annotated Chinese Spring genes against the annotated 'CH Campala *Lr22a*' genes produced hits for 5,210 out of the 6,018 genes whereas 4,656 of the annotated 'CH Campala *Lr22a*' genes produced a BLASTN hit in the annotated 'Chinese Spring' genes. Bi-directional BLAST analysis of the annotated Chinese Spring genes and 'CH Campala *Lr22a*' genes identified a total of 4,097 genes that had each other as the top BLAST hit (i.e. groups of paralogs are not included in this dataset).

A total of 808 out of the annotated 6,018 HC Chinese Spring 2D genes did not produce any BLAST hit (cut-off E-value 10e-^{10}) against the annotated HC 'CH Campala *Lr22a*' genes, whereas 1,227 of the annotated 'CH Campala *Lr22a*' genes did not produce a BLAST hit against the annotated Chinese Spring 2D genes. This would indicate a unique or genotype-specific gene fraction of 13.4% and 20.8% in Chinese Spring and 'CH Campala *Lr22a*', respectively. However, BLAST analysis of these putatively unique genes against the 'CH Campala *Lr22a*' and Chinese Spring 2D pseudomolecules revealed that 782 of the 808 putatively unique Chinese Spring genes and 1,184 of the 1,227 putatively unique 'CH Campala *Lr22a*' genes were present on the pseudomolecule. We randomly selected and validated 20 of the 1,184 putatively unique 'CH Campala *Lr22a*' genes that produced a BLAST hit on the Chinese Spring 2D pseudomolecule and we found intact full-length open reading frames with a 100% sequence identity. Similarly, a random selection of 10 out of the 782 putatively unique Chinese Spring genes revealed that seven genes shared a 100% sequence identity with the respective nucleotide sequence on the 'CH Campala *Lr22a*' 2D pseudomolecule. Hence, these genes were most likely missed or

differentially classified (different confidence classes) by the annotation pipeline. In fact, there were only 26 genes (0.43% of the total genes) that were unique to Chinese Spring (genes that did not show BLAST hit against the annotated genes as well as against the pseudomolecule). Of these, 17 fell into the diverse haploblock *c* on the long chromosome arm and two into haploblock *a* on the short arm of the chromosome. In 'CH Campala *Lr22a*', 43 genes (0.73% of the total genes) were unique of which 14 were from the diverse haploblock *c* and seven from the *Lr22a* introgression region (haploblock *a*). The unique genes in Chinese Spring and 'CH Campala *Lr22a*' are listed in Supplementary table S3.2.

There was a high degree of collinearity with only 169 genes that were non-collinear along the 2D chromosome (e.g. the top BLAST hit of the respective gene was not in the syntenic position in the other genotype) (Supplementary figure S3.3). Of the non-collinear genes, 2, 1, 110 and 11 were from the three diverse haploblocks *a*, *b, c* and *d*, respectively. Since the 'CH Campala *Lr22a*' pseudomolecule was produced by anchoring 'CH Campala *Lr22*a' scaffolds to the IWGSC RefSeq v1.0, we only took into account 'CH Campala *Lr22*a' scaffolds that contained two or more genes for the collinearity analysis.

3.2.5 Chromosome-wide comparison of NLR genes reveals extensive copy number variation in certain NLR families

Regions harboring NLR genes have been reported to be fast evolving to keep up in the arms-race with pathogens (Isidore et al, 2005). Interestingly, three of the four large InDels identified created CNV for NLR genes. We were therefore interested in the dynamics of chromosomal regions harboring NLR genes. For chromosome 2D, a total of 161 NLRs were annotated in the wheat line 'CH Campala *Lr22a*' and 158 NLRs for Chinese Spring. The NLRs annotated in the two wheat genotypes showed a high tendency of clustering and they were mostly located in the telomeric regions (Fig. 3.4a), as it is typically found for this gene class (Internationaal Wheat Genome Sequencing Consortium, 2018).

For ‘CH Campala *Lr22a*’, we found that 62 NLR genes resided in seven gene clusters which comprise 38.5% of the total annotated NLRs. The largest cluster contained 19 NLR genes. In Chinese Spring, we found that 71 NLR genes resided in ten clusters which comprise 44.9% of the total annotated NLRs and the largest cluster contained 21 NLRs. A phylogenetic tree revealed that most NLR genes from Chinese Spring had one ortholog in ‘CH Campala *Lr22a*’ (Fig. 3.4b). On the other hand, we also observed copy number variation for certain regions. Two regions, CNV1 and CNV2, were of particular interest because there was an extensive variation in the NLR copy number between Chinese Spring and ‘CH Campala *Lr22a*’ (Fig. 3.4b). In the CNV1 region, ‘CH Campala *Lr22a*’ had sixteen NLR genes annotated in a 786 kb region. The corresponding region in Chinese Spring contained only two NLRs in a 21 kb interval (Fig. 3.5a). There was a high degree of gene collinearity flanking the NLR cluster (Fig. 3.5a). The two NLR copies in Chinese Spring (NLR46 and NLR47) showed 44% sequence identity at the protein level, indicating that they might have arisen from a very ancient gene duplication. The low sequence identity of NLR46 and NLR47 allowed to assign each of the ‘CH Campala *Lr22a*’ NLRs to one of the two Chinese Spring copies. This revealed a random pattern, which might be explained by complex duplication and rearrangement events (Fig. 3.5a). The CNV1 region locates to the diverse haploblock *c,* which might explain the extent of the CNV found in this region.

The CNV2 region affected a segment of ten paralogous NLR genes situated in a 716 kb region in Chinese Spring. In ‘CH Campala *Lr22a*’ there was a 677 kb deletion that affected all but two of the NLRs. This CNV2 locates in the collinear region between haploblock *c* and haploblock *d*. For this CNV region we could identify a clear breakpoint at one end whereas the other end had a sequence gap (Fig. 3.5b and 3.5c).

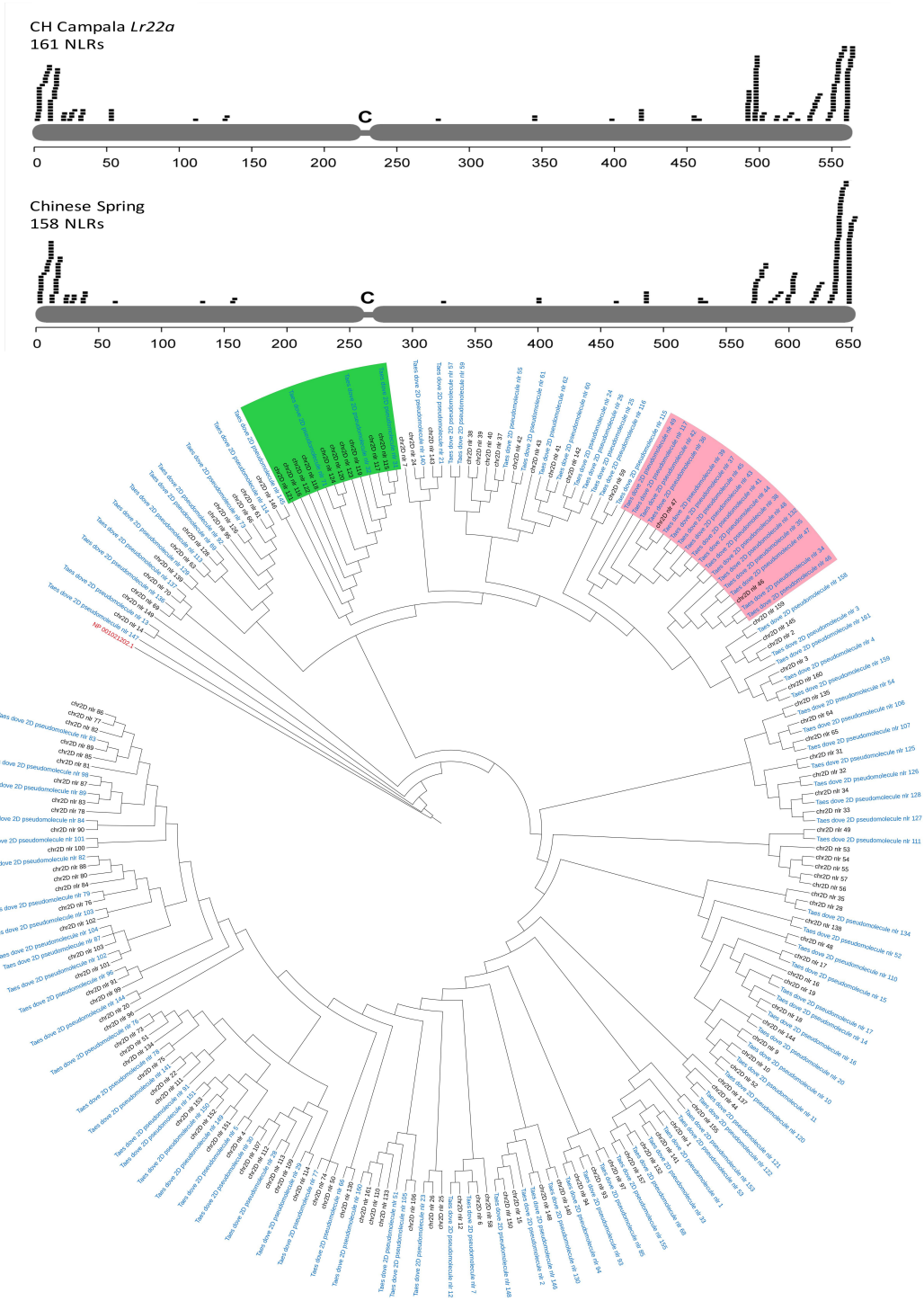


Fig. 3.4 Distribution of predicted NLR genes on chromosome 2D. a The x-axis indicates the position in Mb. Note that the scales differ between 'CH Campala *Lr22a*' and Chinese Spring, because the sequence assembly of 'CH Campala *Lr22a*' is shorter than that of Chinese Spring. **b** Phylogenetic tree where blue labels 'Taes dove 2D pseudomolecule nlr' represent the 'CH Campala *Lr22a*' NLRs and black labels 'chr2D nlr' represent the Chinese Spring NLRs. The two highlighted regions in green and pink represent chromosomal segments with high copy number variation that are discussed in the text.

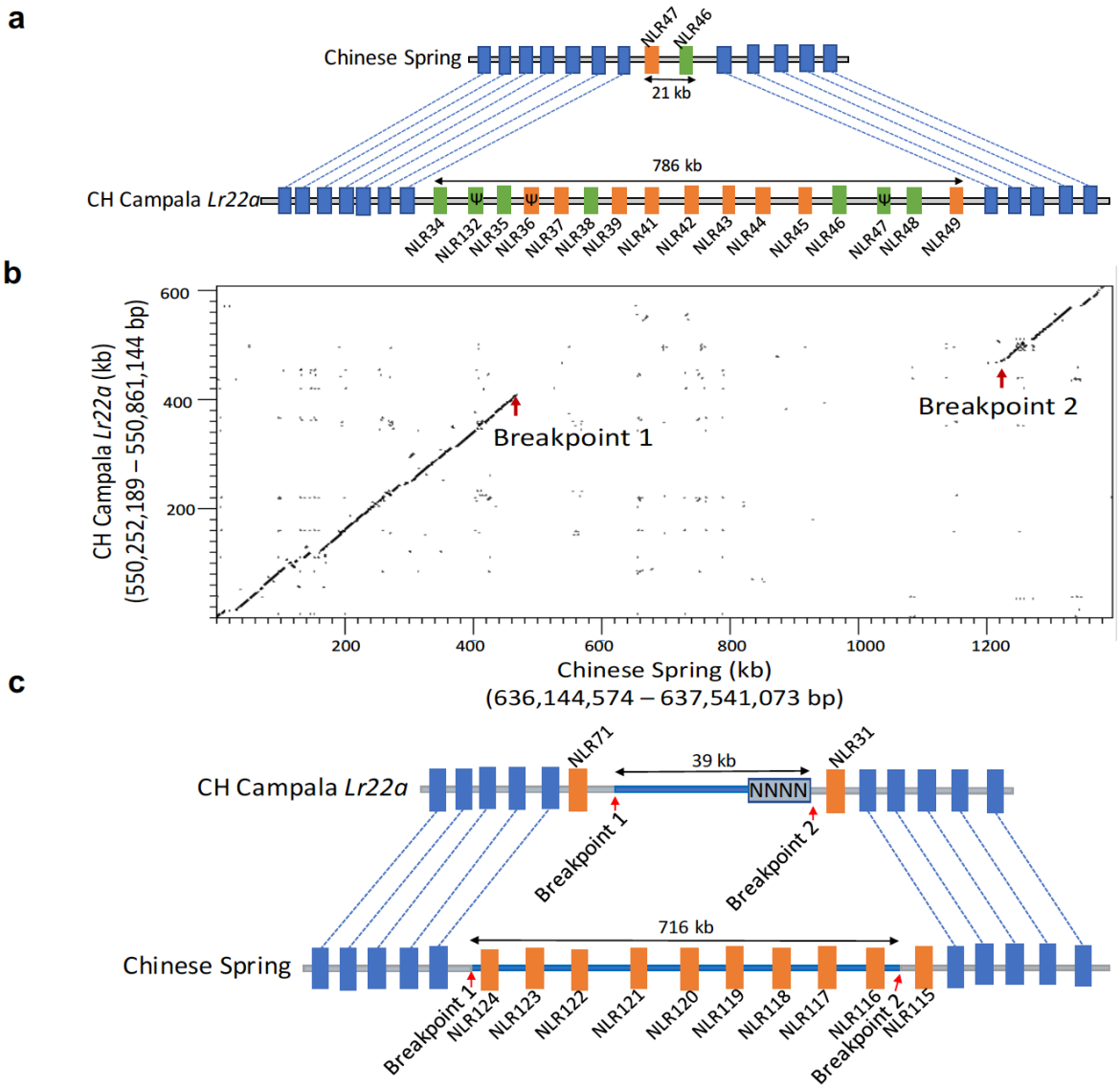


Fig. 3.5 NLR copy number variation. a In the CNV1 region we found 16 NLRs in 'CH Campala *Lr22a*' annotated in a 786 kb region. Pseudogenes are marked with Ψ. Chinese Spring has only two NLRs in a 21 kb segment. **b** NLR gene expansion in Chinese Spring. Dot plot of the CNV region between Chinese Spring and 'CH Campala *Lr22a*'. **c** Chinese Spring had 21 NLRs compared to 14 in 'CH Campala *Lr22a*' which are shown in orange and the collinear genes in the flanking region are shown in blue.

3.3 Discussion

3.3.1 Molecular mechanisms of structural variations

Different genotypes within a plant species can show tremendous genetic diversity. Beside SNPs, SVs have been identified as a major contributor to phenotypic variation in plants, which is why an understanding of large SVs is of importance for breeding (Saxena et al., 2014). For example, the durable fungal stem rust resistance gene *Sr2* of wheat was localized to a region on chromosome 3B that showed extensive structural rearrangements between the *Sr2*-carrying wheat cultivar Hope and the susceptible Chinese Spring on an 867 kb chromosome segment (Mago et al., 2014). How this structural rearrangement affects the *Sr2*-mediated stem rust resistance is not yet understood. Similarly, large deletions comprising multiple tandemly duplicated transcription factor genes at the *Frost resistance-2* locus are associated with reduced frost tolerance in wheat (Pearce et al., 2013). While short-read sequencing allowed a comprehensive assessment of genome-wide SNP distributions in cereals (Chia et al., 2012; The 3,000 rice genomes project, 2014) the identification of SVs, particularly large InDels, has been challenging due to technical limitations. In wheat, the lack of high-quality chromosome assemblies from multiple genotypes has prevented such comparisons so far. Even for other cereal crop species like rice, maize, barley and sorghum there are no or only very few high-quality *de novo* assemblies available beside the reference genotypes (International Rice Genome Sequencing Project, 2005; Mascher et al., 2017; Paterson et al., 2009; Schnable et al., 2009). Here, we compared two high-quality sequence assemblies of bread wheat chromosome 2D that were highly contiguous over megabases, which allowed us to focus on InDels of several hundred kb in size. In total, we found that around 0.3% of the chromosome was affected by the four large InDels. Based on these numbers, we estimate that a comparison of any two wheat genotypes would reveal around 30 large InDels affecting ~15 Mb across the entire D sub-genome. Not surprisingly, the number of small

InDels is much higher than larger structural rearrangements. For example, a comparison of the B73 maize reference assembly to optical maps generated from the two maize inbred lines Ki11 and W22 revealed around 3,400 insertions and deletions between two maize lines with an average InDel size of 20 kb (Jiao et al., 2017). A re-sequencing study in rice revealed a total of 13,045 insertions and 15,151 deletions in the size range of 10-1,000 bp (Vaughn & Bennetzen, 2014). Large InDels affected multiple genes and can therefore have a deleterious effect, particularly in diploid species.

Unequal crossing over and DSB repair were identified as the molecular mechanisms responsible for large InDels in our study. Analyses in Brachypodium revealed that DSB repair is the most common mechanism for structural rearrangements (Buchmann et al., 2012; Wicker et al., 2010). The error prone DSB repair leads to insertions, deletions or rearrangements in the genome. In our comparative analysis, we found a large deletion of 494 kb in Chinese Spring where DSB repair via single strand annealing led to the deletion of the intervening region between the conserved motifs known as DSB signatures. Similar mechanisms were identified in a comparative analysis of the two barley cultivars Barke and Morex, where DSB repair accounted for 41% of the InDel events (Munoz-Amatriain et al., 2013). DSB repair signatures were also found in maize where they flanked small InDels ranging from 5 bp to 175 bp (Woodhouse et al., 2010). Apart from DSB repair, another frequently observed mechanism for SV is unequal crossing over. We found a 285 kb deletion in Chinese Spring where the deletion was a result of an improper alignment of two highly similar NLR genes that served as a template for unequal crossing over. Unequal crossing over has been shown to be one of the main driving forces for genome differences and has been reported to occur in various disease resistance gene families where they result in novel specificities and haplotypes (Cai & Xu, 2007). For example, unequal crossing over between homologs in the maize rust resistance locus *Rp1* led to the formation of recombinant genes with diverse resistance specificities (Ramakrishna et al., 2002; Sudupak et

al., 1993). In soybean, unequal crossing over at the *RPS* locus was associated with loss of resistance to Phytophthora due to the deletion of a NLR-like (*NBSRps4/6)* sequence (Sandhu et al., 2004).

3.3.2 Identification of diverse haploblocks – implications for wheat D-genome dynamics

In addition to SVs, the chromosome-scale assemblies also allowed us to assess SNP density across the entire chromosome and to identify large contiguous blocks with strong variation from the average SNP density. This revealed the presence of four haploblocks that showed a much higher SNP density compared to the rest of the chromosome. One of these haploblocks (haploblock *a*) could be traced back to an artificial introgression that carries the adult plant leaf rust resistance gene *Lr22a* (Hiebert et al., 2007; Thind et al., 2017).*Lr22a* was introgressed into hexaploid wheat by artificially hybridizing the tetraploid wheat line tetra-Canthatch with the diploid *Ae. tauschii* accession RL 5271 (Dyck & Kerber, 1970). The crossing of tetraploid wheat with diverse *Ae. tauschii* accessions results in so called synthetic wheat. This is a widely explored strategy in breeding to compensate for the loss of diversity in hexaploid wheat that went along with domestication and modern breeding (Dreisigacker et al., 2008; Mcfadden & Sears, 1944; Tanksley & McCouch, 1997). After this initial cross, the resulting synthetic hexaploid wheat line was backcrossed six times with the historically important North American wheat cultivar Thatcher, which resulted in the *Lr22a*-containing backcross line 'Thatcher *Lr22a*' (RL 6044). This backcross line then served as the donor to transfer *Lr22a* into elite wheat cultivars including the Canadian wheat cultivar 'AC Minto' and the Swiss spring wheat line 'CH Campala *Lr22a*' (Hiebert et al., 2007; Moullet et al., 2014). The SNP density analysis allowed us now to infer the size of the remaining RL 5271 segment after a limited number of crosses. We did not find evidence for co-introduction of additional segments from the original *Ae. tauschii* donor along chromosome 2D. More interestingly, three additional diverse

haploblocks (haploblocks *b*, *c* and *d*) of almost 9 Mb, 48 Mb and 4 Mb were identified towards the telomeric end of the short and long chromosome arms, respectively. It has been reported that the wheat D genome was most likely contributed by an *Ae. tauschii* population from a region close to the southern or southwestern Caspian Sea. This accession belonged to one of two genetically distinct sublineages within the *Ae. tauschii* gene pool (sublineage 2) (Wang et al., 2013). However, it has been found that gene flow from *Ae. tauschii* accessions belonging to the genetically distant sublineage 1 occurred after the formation of hexaploid wheat, which might explain the presence of contiguous haploblocks with increased diversity. Interestingly, Wang et al. (2013) identified a putative introgression of *Ae. tauschii* sublineage 1 on the telomeric end of chromosome arm 2DL in hexaploid wheat, which might be identical to the diverse haploblock *c* identified in our study. Alternatively, these diverse haploblocks might stem from an alien introgression from another grass species. Interspecies hybridizations are a common method in wheat breeding to transfer specific traits from wild and domesticated grasses into wheat (Molnár-Láng et al., 2015). In contrast to the naturally occurring gene flow from *Ae. tauschii*, the vast majority of these alien introgressions were artificially produced and require *in-vivo* culture techniques like embryo rescue. The length of the haploblock *c* was surprising because the size of haploblocks is expected to be negatively correlated with recombination rates (Greenwood et al., 2004). Since the haploblock *c* located to the highly-recombining telomeric end of the chromosome, we would expect that its size decreases over time. One explanation for conservation of this haploblock could be that its presence suppresses recombination in this area. In contrast to haploblocks *a*, *b* and *d,* we observed a breakdown of sequence homology in intergenic regions in haploblock *c*. On the other hand, the gene order was largely collinear in haploblock *c*, which should be sufficient for recombination in this chromosome segment. A second explanation is that this haploblock *c* might be widely present in the wheat gene pool or in particular breeding programs. For example, PCR analysis revealed that the haploblock *c* was present in

multiple CIMMYT wheat lines. This would allow recombination in the haploblock without decreasing its size. In summary, a considerable fraction of the chromosome (10%) was made up of haploblocks with a much greater diversity than the rest of the chromosome. This highlights the importance of natural gene flow and artificial hybridization as sources for diversity in cereal breeding.

3.3.3 Comparative genomics: Real differences vs. artefacts – a note of caution

In addition to a better understanding of genome dynamics, our analysis also revealed that manual inspection of variation revealed by automated scripts is required in order to distinguish true variants from assembly or annotation artefacts. For example, 22 of the 26 initially identified large InDels had 'Ns' at both ends, indicating that they were most likely due to mis-assembly in one or the other genotype. A similar observation was made for the gene annotation. Our initial comparison of annotated genes revealed a high proportion (14%-21%) of genes that were uniquely present in only one of the two wheat genotypes. Careful validation of the data however revealed that most of these genotype-specific genes produced a BLAST hit at the syntenic position in the other wheat line, indicating that these genes are present but that they were most likely missed or differentially classified (HC and LC confidence classes) by the annotation pipeline. Potential reasons for this observation include artefacts and errors while aligning gene evidences and predicting gene structures, conflicting transcriptome evidences and truncated or incomplete gene models. The actual fraction of unique genes was considerably lower with only 26 and 43 genes that were truly unique in Chinese Spring and 'CH Campala *Lr22a*', respectively. A recent pan-genome study that was based on short-read resequencing of 18 wheat cultivars compared to a medium-quality Chinese Spring assembly reported a total of 128,656 genes in the genome of hexaploid wheat, of which 49,952 (38.8%) were variable (Montenegro et al., 2017). On chromosome 2D, 3.3% - 11% of the 4,703 annotated genes in the respective Chinese

Spring assembly were reported to be absent in the other wheat cultivars. Similarly, Liu et al. (2016) mapped Illumina reads of flow-sorted chromosome 3B of the Fusarium crown rot resistant wheat line CRNIL1A to a high-quality assembly of chromosome 3B from Chinese Spring (Choulet et al., 2014). They identified 499 gene-containing contigs that were specifically found in CRNIL1A but absent in Chinese Spring. The respective Chinese Spring assembly that was used for the comparison contained 5,326 protein-coding genes and hence, the unique gene fraction in CRNIL1A was estimated to be 9.4%. Surprisingly, our conservative approach revealed that the fraction of unique genes is in the range of 0.43% - 0.73% only, which is 5 – 25 fold lower than the estimates that were based on short-read resequencing. It is possible that Chinese Spring and 'CH Campala *Lr22a*' share a particularly high degree of sequence identity on chromosome 2D compared to other cultivars, although there is no obvious connection between the two wheat lines based on the pedigree information. It is therefore more likely that the number of unique genes was overestimated in previous studies, which might have been caused by assembly or annotation artefacts that could not have been accounted for. It has been proposed that the quality of an assembly does affect the quality of gene annotation (Denton et al., 2014). An example for this is the maize line B73, for which two high-quality *de novo* genome assemblies exist. While the first version of the reference sequence predicted 32,540 protein coding genes in the B73 genome (Schnable et al., 2009), a recently released and improved version of the same genotype reported 39,324 protein coding genes (Jiao et al., 2017). The difference of 6,784 genes (17%) can only be explained by technical variation. This example highlights the fact that the assembly quality and annotation procedure can have a tremendous influence on the prediction of the gene content and hence, the estimation of genotype-specific genes. In summary, we provide evidence that the number of unique or variable genes in wheat has been overestimated in past studies due to low assembly qualities and intrinsic variation in genome annotation pipelines.

Hence, the so-called pan-genome of wheat might be considerably smaller than what

was previously estimated (Montenegro et al., 2017). It has to be noted that the wheat D-genome is the least diverse of the three wheat sub-genomes (Akhunov et al., 2010) (Jordan et al., 2015) and it is likely that the fraction of unique genes is higher in the A and B genomes, although most likely not as high as estimated previously. A recent study in *Arabidopsis thaliana* also found that careful manual curation is necessary in order to avoid overestimation of genotype-specific genes. The comparison of high-quality assemblies of the Arabidopsis ecotypes Columbia and Landsberg revealed 63 (0.23%) unique genes in Columbia and 40 (0.14 %) unique genes in Landsberg, which is very similar to the numbers we report in our comparison (Zapata et al., 2016). A comparison of two high-quality assemblies of the *indica* rice lines Zhenshan 97 and Minghui 63 revealed around 4% genotype-specific genes. An important note is that these calculations focused on the presence-absence variation of single genes and did not measure the extent of gene copy number variation as it was for example described for the NLR genes in our study.

3.4 Conclusions

This study provides the first comparison of two wheat pseudomolecules based on high-quality *de novo* chromosome assemblies. The megabase-sized scaffolds allowed us to focus particularly on InDels of several hundred kb in size. Our analysis revealed that around 0.3% of the chromosome was affected by large InDels between the two wheat lines. Our study also revealed that careful manual validation is required in order not to overestimate the frequency of InDels and genotype-specific genes. In particular, 84% of the InDels that were initially identified and 96% of the genotype-specific genes identified through automated pipelines were removed after manual curation because they were most likely due to assembly and annotation artefacts. It is conceivable that previous comparative analyses in wheat that were based on short-read

resequencing alone could not account for these problems. We therefore highlight the importance of manual data validation in future wheat pan-genome projects.

3.5 Methods

3.5.1 'CH Campala *Lr22a*' pseudomolecule assembly

The initial sequence assembly provided by Dovetail Genomics consisted of 10,344 sequence scaffolds (hereafter referred to as Dovetail scaffolds) with and average size of 54.8 kb and an N50 of 9.758 Mb (Thind et al., 2017). To anchor these scaffolds to the IWGSC RefSeq v1.0 chromosome 2D, segments of the scaffolds were used in BLASTN searches against the Chinese Spring chromosome (International Wheat Genome Sequencing Consortium, 2018). Dovetail scaffolds shorter than 10 kb were used in their entirety for the BLASTN search. For Dovetail scaffolds between 10 and 200 kb, a 1 kb segment every 30 kb was used for the BLASTN search. For Dovetail scaffolds larger than 200 kb, a 1 kb segment every 100 kb was used for BLASTN search. For each Dovetail scaffold, it was then determined where the majority of BLAST hits were located in Chinese Spring 2D. Based on this information, Dovetail scaffolds were ordered.

After sequence scaffolds were assembled into a first version of a pseudomolecule, we searched for large-scale breaks in gene collinearity when compared to Chinese Spring chromosome 2D. Here, we focused on blocks of BLASTN hits that mapped to completely different regions of the genome. If the end of a non-collinear block coincided with the end of a Dovetail scaffold, this was interpreted as an assembly artefact. The approximate location of the mis-assembly was identified and the respective Dovetail scaffold was then split into segments. We identified ten putatively chimeric Dovetail scaffolds with assembly errors. These were split into 24 segments (some Dovetail scaffolds contained multiple mis-assemblies) which were then anchored individually to Chinese Spring chromosome 2D. A

total of 7,617 Dovetail scaffolds were integrated to the final pseudomolecule of 563 Mb, representing 73% of all Dovetail scaffolds and 98.92% of the total length of the Dovetail assembly. The integrated 7,617 Dovetail scaffolds have an N50 of 8.78 Mb and an N90 of 1.89 Mb. The scaffold N50 of 8.78 Mb is slightly lower than the N50 of the original assembly obtained from Dovetail Genomics, which is due to the splitting of chimeric scaffolds.

3.5.2 Gene Annotation

We combined two strategies to facilitate gene prediction on the 'CH Campala *Lr22a*' 2D pseudomolecule: prediction using homology from reference proteins and prediction using gene expression data.

For homology-based annotation step, we combined available Triticeae protein sequences obtained from UniProt (05/10/2016), which contain amongst others validated protein sequences from *Triticum aestivum*, *Aegilops tauschii* and *Hordeum vulgare.* These protein sequences were mapped to the nucleotide sequence of the 'CH Campala *Lr22a*' 2D pseudomolecule using the splice-aware alignment software Genomethreader (version 1.6.6, arguments: -startcodon -finalstopcodon -species rice -gcmincoverage 70 -prseedlength 7 -prhdist 4) (Gremme et al., 2005).

In the expression data based step, we used full-length cDNA sequences (leaf, root, seedling, seed, spike and stem (Clavijo et al., 2017) and 1 Full-length cDNA library), as well as multiple RNASeq datasets (E-MTAB-2127, SRP045409, ERP004714/URGI, E-MTAB-21729, PRJEB15048) as evidences to guide the gene structure prediction on the 'CH Campala *Lr22a*' 2D pseudomolecule. Full-length cDNA and IsoSEQ nucleotide sequences were aligned to the pseudomolecule using GMAP (version 2016-06-30, standard parameter, PMID: 15728110), whereas RNASeq datasets were first mapped using Hisat2 (version 2.0.4, parameter: --dta, PMID: 25751142), and subsequently assembled into transcript sequences by

Stringtie (version 1.2.3, parameter: m 150 -t -f 0.3, PMID: 25690850). All transcripts from flcDNA, IsoSeq and RNASeq were combined using Cuffcompare (version 2.2.1, PMID: 26519415) and merged with Stringtie (version 1.2.3, parameter: --merge -m 150) to remove fragments and redundant structures. Next, we used Transdecoder (version 3.0.0) to find potential open reading frames and to predict protein sequences. We used BLASTP (ncbi-blast-2.3.0+, parameter: -max_target_seqs 1 -evalue $1e^{-05}$, PMID: 2231712) to compare potential protein sequences with a trusted protein reference database (Uniprot Magnoliophyta, reviewed/swissprot, downloaded on 03. Aug 2016) and used hmmscan (version 3.1b2, PMID: 22039361) to identify conserved protein family domains for all potential proteins. BLAST and hmmscan results were fed back into Transdecoder-predict to select best translations per transcript sequence.

Finally, all results were combined and redundant protein sequences were removed to form a single non-redundant candidate dataset. In order to differentiate candidates into complete and valid genes, non-coding transcripts, pseudogenes and transposable elements, we applied a confidence classification protocol. Candidate protein sequences were compared against the following 3 manually curated databases using BLAST.

First, “PTREP”, a database of hypothetical proteins contains deduced amino acid sequence, in which in many cases frameshifts were removed. PTREP is useful for the identification of divergent TEs having no significant similarity at the DNA level. Second, UniPoa. a database compromised of annotated poaceae proteins. Third, UniMag, a database of validated magnoliophyta proteins. UniPoa and UniMag protein sequences were downloaded from Uniprot on 30. Aug 2016 and further filtered for complete sequences with start and stop codon. Best hits were selected for each predicted protein to each of the three databases. Only hits with an E-Value below $10e^{-10}$ were considered.

Furthermore, only hits with subject coverage (for protein references) or query coverage (transposon database) above 90 % were considered significant and protein sequences were further classified using the following confidence.

High confidence (HC): Protein sequence is complete and has a subject and query coverage above the threshold in database UniMag (HC1), or no blast hit in database UniMag, but in UniPoa and not TREP (HC2).

Low confidence (LC): Protein sequence is not complete and hit in database UniMag or UniPoa, but not in TREP (LC1), or hit not in UniMag and UniPoa and TREP, but protein sequence is complete.

The tag REP was assigned for protein sequences not in UniMag and complete, but hits in TREP.

In a last step, a set of representative genes within the HC group was selected by choosing the longest transcript for each predicted gene model.

3.5.3 NLR annotation and phylogenetic tree

NLR loci on the 'CH Campala *Lr22a*' pseudomolecule was annotated using NLR-Annotator (https://github/steuernb/NLR-Annotator). The initial fragmentation step of NLR-Annotator was performed generating 20 kb fragments that overlap by 5 kb. Multiple alignments of NB-ARC associated amino acid motifs were generated using NLR-Annotator (output option –a). Multiple alignment files were concatenated and a comparative phylogenetic tree was generated using FastTree (http://www.microbesonline.org/fasttree/) version 2.1.7 (Price et al., 2010).

3.5.4 Identification of the SVs

We analysed SVs in the telomeric and interstitial regions and excluded the centromeric region which was ~100 Mb in size (position 190-290 Mb in Chinese Spring pseudomolecule

and 150-250 Mb in 'CH Campala *Lr22a*' pseudomolecule). The centromeric region is extremely repetitive and gene poor and alignments were difficult. For the identification of the SVs, we segmented the Chinese Spring and 'CH Campala *Lr22a*' pseudomolecules in the windows of 10 Mb and performed dot plot alignments (program DOTTER) (Sonnhammer & Durbin, 1995). For each of the InDels observed, we analysed the sequence alignments to identify the region where the sequence similarity broke down and this region was called breakpoint. We spliced out 5 kb sequence upstream and downstream of these breakpoints and performed BLASTN search (Altschul et al., 1997) against the repeat database to identify transposable elements and also against the *Brachypodium distachyon* coding sequence database (International Brachypodium Initiative, 2010) to identify genes in the flanking regions to understand the molecular mechanism underlying the observed SVs.

To identify NLR CNV, we compared the NLR clusters in Chinese Spring and 'CH Campala *Lr22a*' and identified the breakpoints as described above. The sequences upstream and downstream of breakpoints were used to identify the collinear genes using BLAST search against the annotated 'CH Campala *Lr22a*' and Chinese Spring genes. Putative start and stop codons of the annotated NLRs were identified based on the orthologs of these NLRs in *Brachypodium distachyon.* The coding sequences of these *Brachypodium distachyon* NLRs were taken from the *Brachypodium distachyon* coding sequence database (International Brachypodium Initiative, 2010) and were used for the dot plot alignment to identify the coding sequence of the Chinese Spring and 'CH Campala *Lr22a*' NLRs. Pseudogenes were predicted on the basis of frameshift mutations, premature stop codon or insertion of a transposable element resulting in a pseudogene.

3.5.5 Haploblock analysis and validation

For the identification of the haploblock region, we mapped previously generated Illumina reads of 'CH Campala *Lr22a*' and 'CH Campala' (Thind et al., 2017)

to the Chinese Spring pseudomolecule using the CLC Main Workbench 7 (Qiagen) with standard parameters. The mapped read file was later used for the variant call analysis on the CLC Main Workbench 7 (Qiagen) using standard parameters. SNP density was calculated in sliding windows of 2.5 Mb. To verify the haploblock *c* region we designed a PCR probe (forward primer-GCCACGAGCGTGGTCGTG and reverse primer-CCTTCATAGCTCCGTAGAAG) spanning the left border of the haploblock *c* of 'CH Campala *Lr22a'*. The PCR amplification was performed in 20 µl reaction mixture containing 65 ng of genomic DNA, 1 µl of 2.5 mM dNTP's, 1 µl of 10 µM of each primer and 0.25 units of Sigma Taq polymerase at 60 °C annealing temperature for 35 cycles. The cycling parameters used were, pre-denaturation at 95 °C for 4 min, which was followed by 35 cycles of 95 °C for 30 s, annealing at 60 °C for 30 s, 72 °C for 2 min and a final extension at 72 °C for 10 min. The PCR products were separated on 1.0% agarose gels.

3.6 Declarations

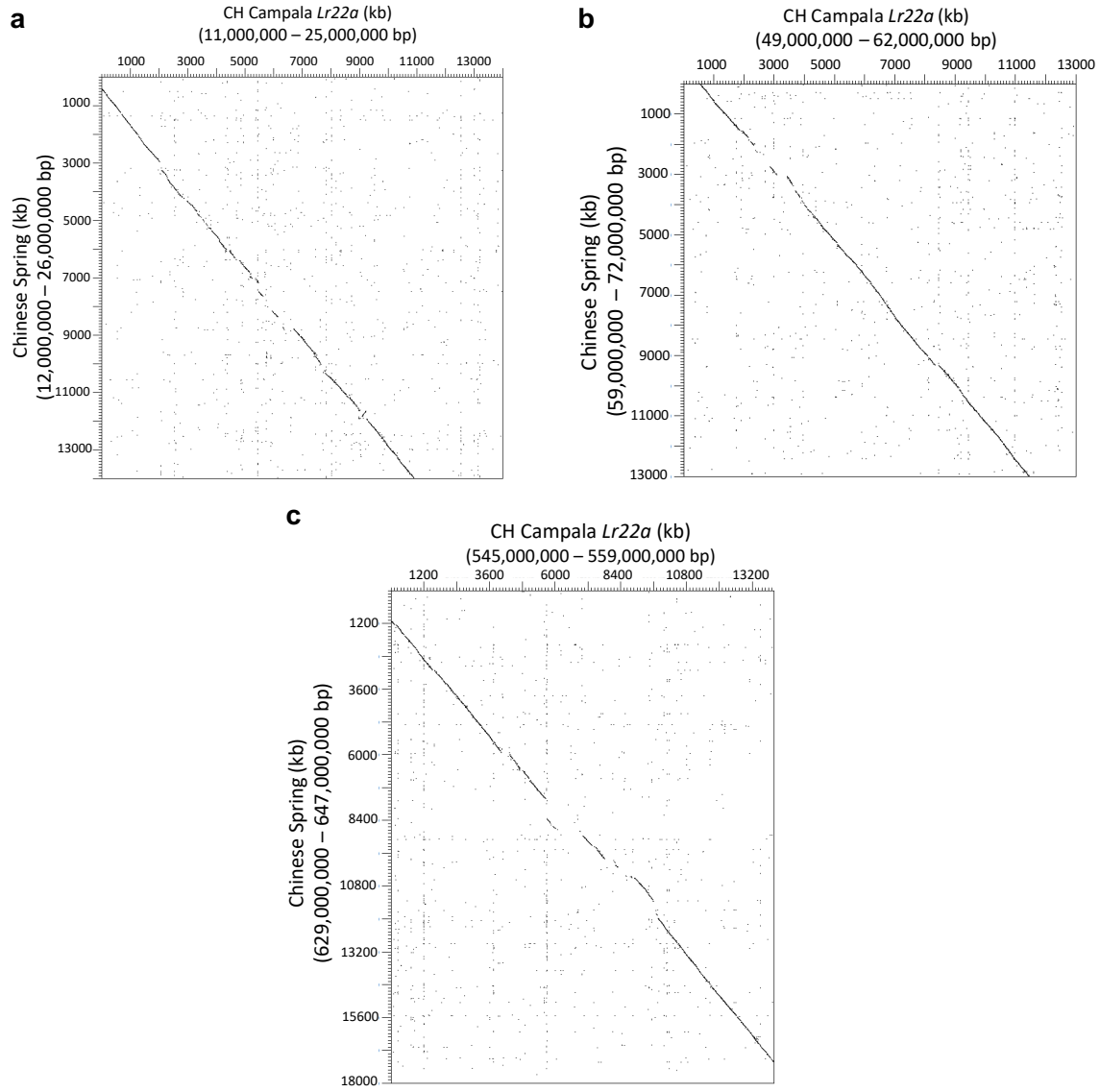
3.6.1 Acknowledgements

We would like to thank Dr. Dario Fossati from Agroscope, Switzerland, Dr. Ravi Singh from CIMMYT and Prof. Jaroslav Dolezel from the Institute of Experimental Botany, Czech Republic for commenting on the possible origin of the large 'CH Camapala *Lr22a*' introgression.

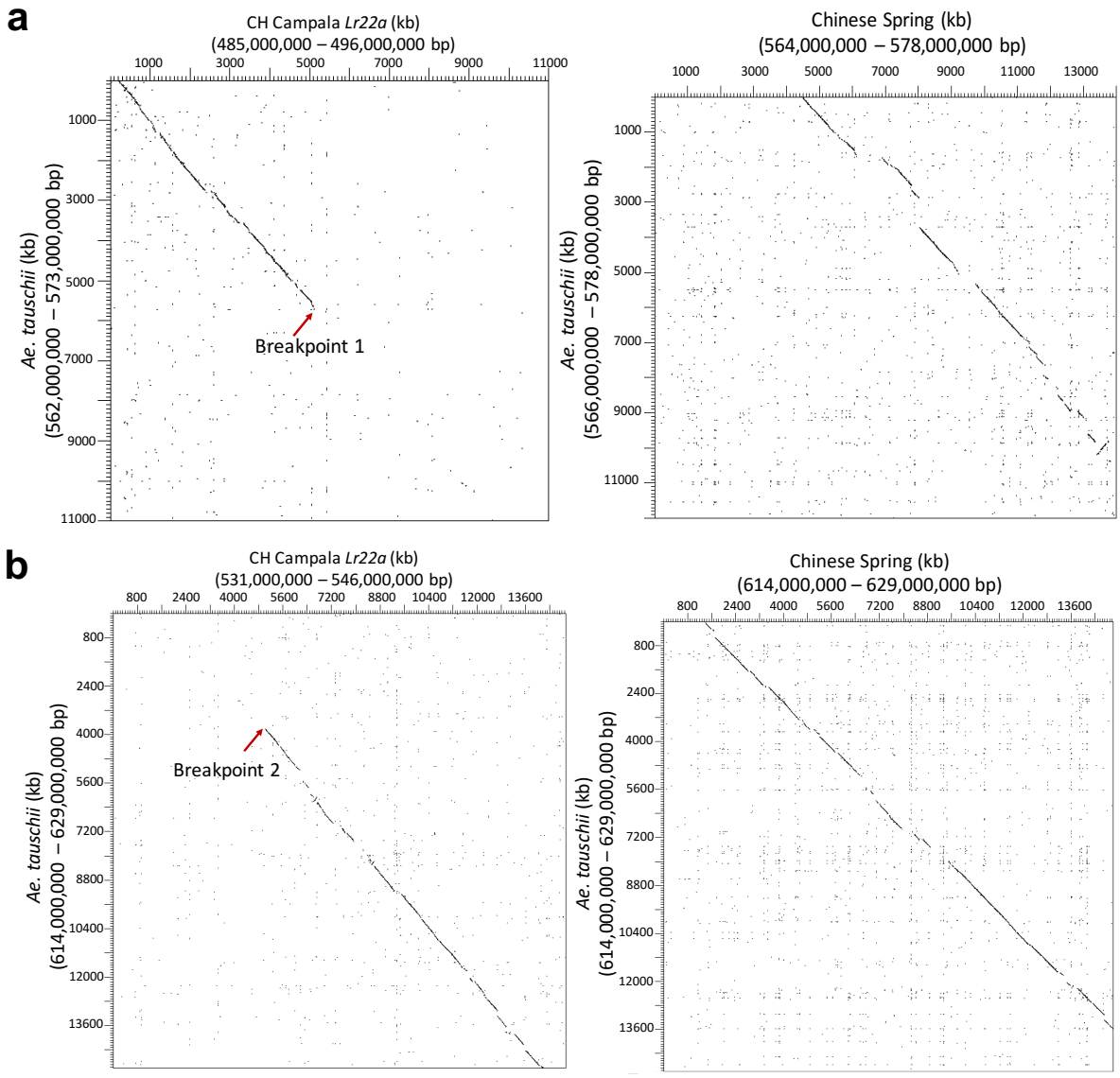
3.6.2 Availability of data and materials

The annotated 'CH Campala *Lr22a*' pseudomolecule was deposited at the European nucleotide archive (ENA) under the accession PRJEB24957.

Supplementary figure S3.1. Haploblocks *a*, *b* and *d* showed sequence homology in the intergenic regions between Chinese Spring and 'CH Campala *Lr22a*'. **a** Dot plot of haploblock *a* with the flanking region represents the *Lr22a* introgression of ~8 Mb in size. **b** Dot plot of ~9 Mb haploblock *b* with the flanking region. **c** Dot plot of the ~4 Mb haploblock *d* with the flanking region. The numbers in brackets refer to the positions of the selected region on the respective pseudomolecule.



Supplementary figure S3.2. Dot plot of the haploblock c region from Chinese Spring and 'CH Campala *Lr22a*' with *Ae. tauschii*. **a** Dot plot of 5 Mb region upstream and 10 Mb downstream of breakpoint 1 of 'CH Campala *Lr22a*' and Chinese Spring with *Ae. tauschii*. **b** Dot plot of 5 Mb region upstream and 10 Mb downstream of breakpoint 2 of 'CH Campala *Lr22a*' and Chinese Spring with *Ae. tauschii*. The numbers in brackets refer to the positions of the selected region on the respective pseudomolecule.



Supplementary table S3.1. List of 678 'CH Campala *Lr22a*' genes that were found in haploblock c.

CH Campala *Lr22a* haploblock c genes	Best Brachipodium hit	Function
TraesCLr22a2Dv5b1G00486500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00486600.1	Bradi5g22080	ABC transporter
TraesCLr22a2Dv5b1G00486700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00486800.1	Bradi5g22090	Remorin, C-terminal region BINDING: protein
TraesCLr22a2Dv5b1G00486900.1	Bradi5g22100	SIT4 phosphatase-associated protein
TraesCLr22a2Dv5b1G00487000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00487100.1	Bradi5g22120	Nucleotide-binding, alpha-beta plait BINDING:, nucleotide, nucleic acid
TraesCLr22a2Dv5b1G00487200.1	Bradi2g35770	Leucine-rich repeat
TraesCLr22a2Dv5b1G00487300.1	Bradi2g35770	Leucine-rich repeat
TraesCLr22a2Dv5b1G00487400.1	Bradi5g19450	GTPase binding
TraesCLr22a2Dv5b1G00487500.1	Bradi5g02370	Leucine-rich repeat
TraesCLr22a2Dv5b1G00487600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00487700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00487800.1	Bradi2g40510	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00487900.1	Bradi4g42470	Cyclin-like F-box
TraesCLr22a2Dv5b1G00488000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00488100.1	Bradi4g42470	Cyclin-like F-box
TraesCLr22a2Dv5b1G00488200.1	Bradi3g19970	NB-ARC
TraesCLr22a2Dv5b1G00488300.1	Bradi1g70070	Transcriptional factor
TraesCLr22a2Dv5b1G00488400.1	Bradi3g16300	Alpha/beta hydrolase
TraesCLr22a2Dv5b1G00488500.1	Bradi4g13290	Protein of unknown function DUF716
TraesCLr22a2Dv5b1G00488600.1	Bradi5g22550	Leucine-rich repeat
TraesCLr22a2Dv5b1G00488700.1	Bradi5g22130	Unknown function
TraesCLr22a2Dv5b1G00488800.1	Bradi5g22200	Bromodomain ACTIVITY: transcription activator
TraesCLr22a2Dv5b1G00488900.1	Bradi5g19980	Protein of unknown function DUF597
TraesCLr22a2Dv5b1G00489000.1	Bradi3g16300	Alpha/beta hydrolase
TraesCLr22a2Dv5b1G00489100.1	Bradi2g18450	Sulfotransferase ACTIVITY: estrone, sulfotransferase BINDING: nucleotide
TraesCLr22a2Dv5b1G00489200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00489300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00489400.1	Bradi1g60260	Alkyl hydroperoxide reductase
TraesCLr22a2Dv5b1G00489500.1	Bradi5g15800	EGF-type Asp/Asn hydroxylation conserved site, ACTIVITY: protein Tyr kinase
TraesCLr22a2Dv5b1G00489600.1	Bradi3g60870	Basic-leucine zipper (bZIP) transcription factor
TraesCLr22a2Dv5b1G00489700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00489800.1	Bradi5g22220	Amine oxidase
TraesCLr22a2Dv5b1G00489900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00490000.1	Bradi3g45080	Pectinesterase inhibitor ACTIVITY: enzyme, inhibitor, pectinesterase

TraesCLr22a2Dv5b1G00490100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00490200.1	Bradi5g22250	FMN-dependent alpha-hydroxy acid dehydrogenase,, active site ACTIVITY: L-lactate
TraesCLr22a2Dv5b1G00490300.1	Bradi5g22290	Glycine cleavage T-protein
TraesCLr22a2Dv5b1G00490400.1	Bradi2g40970	Zinc finger
TraesCLr22a2Dv5b1G00490500.1	Bradi5g22300	Twin-arginine translocation pathway signal
TraesCLr22a2Dv5b1G00490600.1	Bradi0098s00200	Light chain 3 (LC3) BINDING:, phosphatidylethanolamine, microtubule, beta-tubulin
TraesCLr22a2Dv5b1G00490700.1	Bradi3g20540	Pectin lyase fold/virulence factor
TraesCLr22a2Dv5b1G00490800.1	Bradi5g01050	Protein of unknown function DUF594
TraesCLr22a2Dv5b1G00490900.1	Bradi5g22310	Glycogen/starch synthases, ADP-glucose type, ACTIVITY: UDP-glycosyltransferase
TraesCLr22a2Dv5b1G00491000.1	Bradi3g20540	Pectin lyase fold/virulence factor
TraesCLr22a2Dv5b1G00491100.1	Bradi5g22330	ATPase
TraesCLr22a2Dv5b1G00491200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00491300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00491400.1	Bradi5g22340	CDP-diacylglycerol-inositol
TraesCLr22a2Dv5b1G00491500.1	Bradi5g22350	Kelch related ACTIVITY: transcription, regulator BINDING: identical protein
TraesCLr22a2Dv5b1G00491600.1	Bradi3g05480	Leucine-rich repeat
TraesCLr22a2Dv5b1G00491700.1	Bradi5g19360	Histone core BINDING: DNA, protein
TraesCLr22a2Dv5b1G00491800.1	Bradi2g55630	Glycoside hydrolase
TraesCLr22a2Dv5b1G00491900.1	Bradi3g26990	IQ calmodulin-binding region
TraesCLr22a2Dv5b1G00492000.1	Bradi5g22370	Kelch related ACTIVITY: transcription, regulator BINDING: protein
TraesCLr22a2Dv5b1G00492100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00492200.1	Bradi1g45220	Transcription factor TCP subgroup
TraesCLr22a2Dv5b1G00492300.1	Bradi3g60720	GHMP kinase, ATP-binding, conserved site, ACTIVITY: homoSer kinase BINDING: ATP
TraesCLr22a2Dv5b1G00492400.1	Bradi3g39130	F-box associated type 1
TraesCLr22a2Dv5b1G00492500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00492600.1	Bradi3g38380	Uncharacterised protein family UPF0029,, N-terminal BINDING: protein
TraesCLr22a2Dv5b1G00492700.1	Bradi5g22370	Kelch related ACTIVITY: transcription, regulator BINDING: protein
TraesCLr22a2Dv5b1G00492800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00492900.1	Bradi5g22410	Nucleotide-binding, alpha-beta plait
TraesCLr22a2Dv5b1G00493000.1	Bradi5g22420	Ataxin-2
TraesCLr22a2Dv5b1G00493100.1	Bradi5g10840	Zinc finger
TraesCLr22a2Dv5b1G00493200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00493300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00493400.1	Bradi5g22430	HMG-I and HMG-Y, DNA-binding, conserved site
TraesCLr22a2Dv5b1G00493500.1	Bradi5g22350	Kelch related ACTIVITY: transcription, regulator BINDING: identical protein
TraesCLr22a2Dv5b1G00493600.1	Bradi5g22510	Lipocalin-related protein and Bos/Can/Equ, allergen
TraesCLr22a2Dv5b1G00493700.1	Bradi2g35770	Leucine-rich repeat
TraesCLr22a2Dv5b1G00493800.1	Bradi5g22840	Leucine-rich repeat
TraesCLr22a2Dv5b1G00493900.1	Bradi5g22840	Leucine-rich repeat

TraesCLr22a2Dv5b1G00494000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00494100.1	Bradi5g22840	Leucine-rich repeat
TraesCLr22a2Dv5b1G00494200.1	Bradi5g22550	Leucine-rich repeat
TraesCLr22a2Dv5b1G00494300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00494400.1	Bradi5g22550	Leucine-rich repeat
TraesCLr22a2Dv5b1G00494500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00494600.1	Bradi4g09780	Phospholipase
TraesCLr22a2Dv5b1G00494700.1	Bradi2g35770	Leucine-rich repeat
TraesCLr22a2Dv5b1G00494800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00494900.1	Bradi5g22550	Leucine-rich repeat
TraesCLr22a2Dv5b1G00495000.1	Bradi5g22570	Lipid-binding START ACTIVITY: transcription, repressor, transcription factor
TraesCLr22a2Dv5b1G00495100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00495200.1	Bradi5g22580	ABC transporter
TraesCLr22a2Dv5b1G00495300.1	Bradi5g22580	ABC transporter
TraesCLr22a2Dv5b1G00495400.1	Bradi3g15340	PAK-box/P21-Rho-binding
TraesCLr22a2Dv5b1G00495500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00495600.1	Bradi3g05470	Leucine-rich repeat
TraesCLr22a2Dv5b1G00495700.1	Bradi5g22610	Eukaryotic translation initiation factor SUI1
TraesCLr22a2Dv5b1G00495800.1	Bradi5g18030	UDP-glucuronosyl/UDP-glucosyltransferase
TraesCLr22a2Dv5b1G00495900.1	Bradi2g42990	MtN3 and saliva related transmembrane protein
TraesCLr22a2Dv5b1G00496000.1	Bradi5g22630	Protein of unknown function DUF794
TraesCLr22a2Dv5b1G00496100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00496200.1	Bradi4g16240	Leucine-rich repeat
TraesCLr22a2Dv5b1G00496300.1	Bradi1g02140	GrpE nucleotide exchange factor
TraesCLr22a2Dv5b1G00496400.1	Bradi4g38540	Ribosomal protein L40e
TraesCLr22a2Dv5b1G00496500.1	Bradi3g32400	Cyclin-like F-box
TraesCLr22a2Dv5b1G00496600.1	Bradi1g02140	GrpE nucleotide exchange factor
TraesCLr22a2Dv5b1G00496700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00496800.1	Bradi3g32400	Cyclin-like F-box
TraesCLr22a2Dv5b1G00496900.1	Bradi3g09430	Zinc finger
TraesCLr22a2Dv5b1G00497000.1	Bradi3g10720	ATPase
TraesCLr22a2Dv5b1G00497100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00497200.1	Bradi5g22640	Pentatricopeptide repeat
TraesCLr22a2Dv5b1G00497300.1	Bradi5g22650	Haem peroxidase, plant/fungal/bacterial, ACTIVITY: electron carrier, peroxidase
TraesCLr22a2Dv5b1G00497400.1	Bradi1g64920	Bet v I allergen
TraesCLr22a2Dv5b1G00497500.1	Bradi1g64920	Bet v I allergen
TraesCLr22a2Dv5b1G00497600.1	Bradi3g21160	Alpha/beta hydrolase fold-1 activity
TraesCLr22a2Dv5b1G00497700.1	Bradi4g39160	Lipase
TraesCLr22a2Dv5b1G00497800.1	Bradi2g53450	Pentatricopeptide repeat
TraesCLr22a2Dv5b1G00497900.1	Bradi5g00500	Pre-SET zinc-binding sub-group
TraesCLr22a2Dv5b1G00498000.1	Bradi1g54940	Methylmalonate-semialdehyde dehydrogenase
TraesCLr22a2Dv5b1G00498100.1	Bradi4g44400	Plant lipid transfer protein

TraesCLr22a2Dv5b1G00498200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00498300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00498400.1	Bradi3g02130	Terpenoid cylases/protein prenyltransferase
TraesCLr22a2Dv5b1G00498500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00498600.1	Bradi5g22880	Leucine-rich repeat, N-terminal
TraesCLr22a2Dv5b1G00498700.1	Bradi2g23720	Glycoside hydrolase
TraesCLr22a2Dv5b1G00498800.1	Bradi3g46640	Major facilitator superfamily, general substrate, transporter ACTIVITY: mannose
TraesCLr22a2Dv5b1G00498900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00499000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00499200.1	Bradi3g09190	UDP-glucuronosyl/UDP-glucosyltransferase
TraesCLr22a2Dv5b1G00499100.1	Bradi3g48970	Pre-SET zinc-binding sub-group
TraesCLr22a2Dv5b1G00499300.1	Bradi1g29140	Phosphatidylinositol-4-phosphate 5-kinase
TraesCLr22a2Dv5b1G00499400.1	Bradi5g22710	Cyclin-related 2 ACTIVITY: cyclin-dependent, protein kinase regulator
TraesCLr22a2Dv5b1G00499500.1	Bradi5g22710	Cyclin-related 2 ACTIVITY: cyclin-dependent, protein kinase regulator
TraesCLr22a2Dv5b1G00499600.1	Bradi4g07540	Legume lectin, beta domain
TraesCLr22a2Dv5b1G00499700.1	Bradi5g25370	Protein of unknown function DUF724
TraesCLr22a2Dv5b1G00499800.1	Bradi3g16030	DNA polymerase delta
TraesCLr22a2Dv5b1G00499900.1	Bradi5g22730	Zinc finger
TraesCLr22a2Dv5b1G00500000.1	Bradi4g45140	Phosphopantetheine attachment site
TraesCLr22a2Dv5b1G00500100.1	Bradi5g22750	Helicase
TraesCLr22a2Dv5b1G00500200.1	Bradi2g52470	Peptidase C14
TraesCLr22a2Dv5b1G00500300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00500400.1	Bradi4g21850	Leucine-rich repeat
TraesCLr22a2Dv5b1G00500500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00500600.1	Bradi3g23300	Nitrogen regulatory PII-like, alpha/beta, BINDING: enzyme
TraesCLr22a2Dv5b1G00500700.1	Bradi5g22770	Thioredoxin-like fold ACTIVITY: thioredoxin-disulfide reductase
TraesCLr22a2Dv5b1G00500800.1	Bradi3g05480	Leucine-rich repeat
TraesCLr22a2Dv5b1G00500900.1	Bradi5g22780	ATPase BINDING: microtubule, protein kinase, ATP
TraesCLr22a2Dv5b1G00501000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00501100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00501300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00501400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00501500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00501600.1	Bradi4g31620	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00501700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00501800.1	Bradi3g00830	Mitochodrial transcription termination
TraesCLr22a2Dv5b1G00501900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00502000.1	Bradi2g27970	Protein phosphatase 2C
TraesCLr22a2Dv5b1G00502100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00502200.1	Bradi1g06560	Glycosyltransferase AER61
TraesCLr22a2Dv5b1G00502300.1	No homolog	No homology

TraesCLr22a2Dv5b1G00502400.1	Bradi4g38260	Nucleotide-binding, alpha-beta plait ACTIVITY:, RNA transmembrane transporter
TraesCLr22a2Dv5b1G00502500.1	Bradi1g47620	Leucine-rich repeat
TraesCLr22a2Dv5b1G00502600.1	Bradi4g06970	Leucine-rich repeat
TraesCLr22a2Dv5b1G00502700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00502800.1	Bradi3g21160	Alpha/beta hydrolase fold-1 ACTIVITY:, carboxylesterase
TraesCLr22a2Dv5b1G00502900.1	Bradi5g22820	NAD-dependent epimerase/dehydratase
TraesCLr22a2Dv5b1G00503000.1	Bradi5g22830	NAD-dependent epimerase/dehydratase
TraesCLr22a2Dv5b1G00503100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00503200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00503300.1	Bradi5g22830	NAD-dependent epimerase/dehydratase
TraesCLr22a2Dv5b1G00503500.1	Bradi2g23450	Uncharacterised protein family UPF0497
TraesCLr22a2Dv5b1G00503600.1	Bradi5g22830	NAD-dependent epimerase/dehydratase
TraesCLr22a2Dv5b1G00503400.1	Bradi5g22830	NAD-dependent epimerase/dehydratase
TraesCLr22a2Dv5b1G00503700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00503800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00503900.1	Bradi5g13140	Nucleotide-binding, alpha-beta plait BINDING:, nucleotide, protein, poly(A)
TraesCLr22a2Dv5b1G00504000.1	Bradi3g49500	Surfeit locus 6 ACTIVITY:, nucleoside-triphosphatase
TraesCLr22a2Dv5b1G00504100.1	Bradi5g01370	BINDING: protein
TraesCLr22a2Dv5b1G00504200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00504300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00504400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00504500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00504600.1	Bradi4g26560	UDP-glucuronosyl/UDP-glucosyltransferase
TraesCLr22a2Dv5b1G00504700.1	Bradi1g53290	Harpin-induced 1
TraesCLr22a2Dv5b1G00504800.1	Bradi3g45810	N-6 adenine-specific DNA methylase
TraesCLr22a2Dv5b1G00504900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00505000.1	Bradi4g26560	UDP-glucuronosyl/UDP-glucosyltransferase
TraesCLr22a2Dv5b1G00505100.1	Bradi5g22820	NAD-dependent epimerase/dehydratase
TraesCLr22a2Dv5b1G00505200.1	Bradi2g43150	Plant PDR ABC transporter associated
TraesCLr22a2Dv5b1G00505300.1	Bradi5g22900	Major facilitator superfamily,norepine ransmembrane
TraesCLr22a2Dv5b1G00505400.1	Bradi1g06860	Translation elongation factor EF1A/initiation, factor IF2gamma, C-terminal ACTIV
TraesCLr22a2Dv5b1G00505500.1	Bradi5g22900	Major facilitator superfamily,norepine ransmembrane
TraesCLr22a2Dv5b1G00505600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00505700.1	Bradi5g22900	Major facilitator superfamily,norepine ransmembrane
TraesCLr22a2Dv5b1G00505800.1	Bradi5g22900	Major facilitator superfamily,norepine ransmembrane
TraesCLr22a2Dv5b1G00505900.1	Bradi5g22910	Concanavalin A-like lectin/glucanase
TraesCLr22a2Dv5b1G00506000.1	Bradi5g08250	Oligopeptide transporter OPT superfamily
TraesCLr22a2Dv5b1G00506100.1	Bradi5g08250	Oligopeptide transporter OPT superfamily
TraesCLr22a2Dv5b1G00506200.1	Bradi1g62860	Dormancyauxin associated
TraesCLr22a2Dv5b1G00506300.1	Bradi5g22920	Helix-loop-helix DNA-binding
TraesCLr22a2Dv5b1G00506400.1	Bradi5g22930	Serine/Thr protein kinase

TraesCLr22a2Dv5b1G00506500.1	Bradi5g23060	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00506600.1	Bradi5g23060	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00506700.1	Bradi5g23060	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00506800.1	Bradi5g23060	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00506900.1	Bradi2g12880	Phospholipid/glycerol acyltransferase
TraesCLr22a2Dv5b1G00507000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00507100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00507200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00507400.1	Bradi5g23060	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00507500.1	Bradi5g23080	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00507600.1	Bradi5g19700	Pectinesterase inhibitor
TraesCLr22a2Dv5b1G00507700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00507800.1	Bradi1g56830	Alpha/beta hydrolase
TraesCLr22a2Dv5b1G00507900.1	Bradi2g11690	Plant specific eukaryotic initiation factor 4B
TraesCLr22a2Dv5b1G00508000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00508100.1	Bradi1g54210	Glycoside hydrolase
TraesCLr22a2Dv5b1G00508200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00508300.1	Bradi5g23090	Curculin-like (mannose-binding) lectin
TraesCLr22a2Dv5b1G00508400.1	Bradi1g54210	Glycoside hydrolase
TraesCLr22a2Dv5b1G00508500.1	Bradi5g23110	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00508600.1	Bradi5g25410	2OG-Fe(II) oxygenase
TraesCLr22a2Dv5b1G00508700.1	Bradi5g25400	Phospholipid/glycerol acyltransferase
TraesCLr22a2Dv5b1G00508800.1	Bradi5g23130	Diacylglycerol kinase
TraesCLr22a2Dv5b1G00508900.1	Bradi3g02240	Leucine-rich repeat
TraesCLr22a2Dv5b1G00509000.1	Bradi1g43170	Short-chain dehydrogenase/reductase
TraesCLr22a2Dv5b1G00509100.1	Bradi5g23110	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00509200.1	Bradi5g23110	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00509300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00509400.1	Bradi5g27400	Transcriptional factor B3
TraesCLr22a2Dv5b1G00509500.1	Bradi5g25390	Unconventional myosin/plant kinesin-like
TraesCLr22a2Dv5b1G00509600.1	Bradi4g17100	Cyclin-like F-box
TraesCLr22a2Dv5b1G00509700.1	Bradi5g25370	Protein of unknown function DUF724
TraesCLr22a2Dv5b1G00509800.1	Bradi5g25340	Zinc finger
TraesCLr22a2Dv5b1G00509900.1	Bradi5g25340	Zinc finger
TraesCLr22a2Dv5b1G00510000.1	Bradi5g25320	Myb-like DNA-binding region
TraesCLr22a2Dv5b1G00510100.1	Bradi5g25320	Myb-like DNA-binding region
TraesCLr22a2Dv5b1G00510200.1	Bradi5g25310	RNA polymerase I, transcription factor
TraesCLr22a2Dv5b1G00510300.1	Bradi3g00830	Mitochodrial transcription termination
TraesCLr22a2Dv5b1G00510400.1	Bradi4g07850	Aminoacyl-tRNA synthetase
TraesCLr22a2Dv5b1G00510500.1	Bradi5g14680	Major facilitator superfamily
TraesCLr22a2Dv5b1G00510600.1	Bradi5g25280	unknown function
TraesCLr22a2Dv5b1G00510700.1	Bradi5g25270	Glycosyl hydrolases
TraesCLr22a2Dv5b1G00510800.1	Bradi5g25250	Leucine-rich repeat

TraesCLr22a2Dv5b1G00510900.1	Bradi5g25200	Leucine-rich repeat
TraesCLr22a2Dv5b1G00511000.1	Bradi5g25250	Leucine-rich repeat
TraesCLr22a2Dv5b1G00511100.1	Bradi3g35170	Harpin-induced 1
TraesCLr22a2Dv5b1G00511200.1	Bradi3g43600	Cytochrome P450
TraesCLr22a2Dv5b1G00511300.1	Bradi2g30790	DNA-binding WRKY
TraesCLr22a2Dv5b1G00511400.1	Bradi3g47790	Phosphopantetheine attachment site
TraesCLr22a2Dv5b1G00511500.1	Bradi5g25200	Leucine-rich repeat
TraesCLr22a2Dv5b1G00511600.1	Bradi4g19460	S-adenosyl-L-homocysteine hydrolase
TraesCLr22a2Dv5b1G00511700.1	Bradi1g28000	Peptidase S10
TraesCLr22a2Dv5b1G00511800.1	Bradi1g47610	Leucine-rich repeat
TraesCLr22a2Dv5b1G00511900.1	Bradi1g02140	GrpE nucleotide exchange factor
TraesCLr22a2Dv5b1G00512000.1	Bradi5g25190	Leucine-rich repeat
TraesCLr22a2Dv5b1G00512100.1	Bradi5g25170	Transferase
TraesCLr22a2Dv5b1G00512200.1	Bradi3g30590	Cytochrome P450
TraesCLr22a2Dv5b1G00512300.1	Bradi3g16110	2OG-Fe(II) oxygenase
TraesCLr22a2Dv5b1G00512400.1	Bradi1g02960	WD40/YVTN repeat-like
TraesCLr22a2Dv5b1G00512500.1	Bradi1g02950	Zinc finger
TraesCLr22a2Dv5b1G00512600.1	Bradi3g35980	Cytochrome P450
TraesCLr22a2Dv5b1G00512700.1	Bradi5g25170	Transferase
TraesCLr22a2Dv5b1G00512800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00512900.1	Bradi5g25160	Transcriptional factor B3
TraesCLr22a2Dv5b1G00513000.1	Bradi5g01720	Peptidyl-prolyl cis-trans isomerase
TraesCLr22a2Dv5b1G00513100.1	Bradi5g25130	WD40 repeat
TraesCLr22a2Dv5b1G00513200.1	Bradi5g22950	Aldehyde dehydrogenase
TraesCLr22a2Dv5b1G00513300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00513400.1	Bradi5g22950	Aldehyde dehydrogenase
TraesCLr22a2Dv5b1G00513500.1	Bradi1g36270	SFT2-like
TraesCLr22a2Dv5b1G00513600.1	Bradi5g25110	PHF5-like ACTIVITY: RNA splicing factor, transesterification mechanism
TraesCLr22a2Dv5b1G00513700.1	Bradi5g25100	Zinc finger
TraesCLr22a2Dv5b1G00513800.1	Bradi1g54770	F-box associated type 1
TraesCLr22a2Dv5b1G00513900.1	Bradi5g25090	IQ calmodulin-binding region
TraesCLr22a2Dv5b1G00514000.1	Bradi5g25080	Regulator of chromosome, condensation/beta-lactamase-inhibitor protein II
TraesCLr22a2Dv5b1G00514100.1	Bradi3g08610	Uncharacterised protein family UPF0054
TraesCLr22a2Dv5b1G00514200.1	Bradi3g22870	ATPase
TraesCLr22a2Dv5b1G00514300.1	Bradi5g25070	GTP cyclohydrolase I
TraesCLr22a2Dv5b1G00514400.1	Bradi2g46250	Glycosyl transferase
TraesCLr22a2Dv5b1G00514500.1	Bradi3g35230	FAD linked oxidase
TraesCLr22a2Dv5b1G00514600.1	Bradi5g25050	2OG-Fe(II) oxygenase
TraesCLr22a2Dv5b1G00514700.1	Bradi4g19460	S-adenosyl-L-homocysteine hydrolase
TraesCLr22a2Dv5b1G00514800.1	Bradi4g19470	Pyridoxal phosphate-dependent transferase
TraesCLr22a2Dv5b1G00514900.1	Bradi5g19450	GTPase BINDING: protein
TraesCLr22a2Dv5b1G00515000.1	Bradi1g60720	Serine/Thr protein kinase

TraesCLr22a2Dv5b1G00515100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00515200.1	Bradi5g25010	Auxin responsive SAUR protein
TraesCLr22a2Dv5b1G00515300.1	Bradi5g25020	Auxin responsive SAUR protein
TraesCLr22a2Dv5b1G00515400.1	Bradi5g25020	Auxin responsive SAUR protein
TraesCLr22a2Dv5b1G00515500.1	Bradi5g25000	Auxin responsive SAUR protein
TraesCLr22a2Dv5b1G00515600.1	Bradi2g43800	Region of unknown function
TraesCLr22a2Dv5b1G00515700.1	Bradi1g51820	Transcription regulator
TraesCLr22a2Dv5b1G00515800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00515900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00516000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00516100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00516200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00516300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00516400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00516500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00516600.1	Bradi5g24980	Ribosomal protein L40e
TraesCLr22a2Dv5b1G00516700.1	Bradi4g10040	Leucine-rich repeat
TraesCLr22a2Dv5b1G00516800.1	Bradi3g56370	unknown function
TraesCLr22a2Dv5b1G00516900.1	Bradi5g24960	Endopeptidase, translation factor
TraesCLr22a2Dv5b1G00517000.1	Bradi2g08140	ACTIVITY: transporter
TraesCLr22a2Dv5b1G00517100.1	Bradi5g24950	MED7 ACTIVITY: RNA polymerase II, transcription mediator
TraesCLr22a2Dv5b1G00517200.1	Bradi4g24410	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00517300.1	Bradi5g24940	ACTIVITY: catalytic BINDING: DNA, protein
TraesCLr22a2Dv5b1G00517400.1	Bradi5g13810	Protein phosphatase
TraesCLr22a2Dv5b1G00517500.1	Bradi5g02740	Pentatricopeptide repeat
TraesCLr22a2Dv5b1G00517600.1	Bradi1g56800	Prefoldin alpha-like BINDING: actin filament,, microtubule, unfolded protein
TraesCLr22a2Dv5b1G00517700.1	Bradi2g38810	Leucine-rich repeat
TraesCLr22a2Dv5b1G00517800.1	Bradi4g11380	Short-chain dehydrogenase/reductase
TraesCLr22a2Dv5b1G00517900.1	Bradi5g24930	Molybdenum cofactor biosynthesis
TraesCLr22a2Dv5b1G00518000.1	Bradi5g24900	HMG-I and HMG-Y, DNA-binding, conserved site
TraesCLr22a2Dv5b1G00518100.1	Bradi5g24890	Inositol-pentakisphosphate 2-kinase
TraesCLr22a2Dv5b1G00518200.1	Bradi5g24880	Heavy metal transport/detoxification protein
TraesCLr22a2Dv5b1G00518300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00518400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00518500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00518600.1	Bradi5g24870	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00518700.1	Bradi2g13840	Heat shock protein
TraesCLr22a2Dv5b1G00518800.1	Bradi4g24650	ABA/WDS induced protein
TraesCLr22a2Dv5b1G00518900.1	Bradi2g04680	Tetratricopeptide-like helical ACTIVITY:, protein kinase regulator
TraesCLr22a2Dv5b1G00519000.1	Bradi3g10310	Concanavalin A-like lectin/glucanase
TraesCLr22a2Dv5b1G00519100.1	Bradi5g24850	Alpha-1,4-glucan-protein synthase
TraesCLr22a2Dv5b1G00519200.1	Bradi5g24840	Protein kinase-like

TraesCLr22a2Dv5b1G00519300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00519400.1	Bradi5g24820	Probable translation factor pelota ACTIVITY:, endoribonuclease BINDING: protein
TraesCLr22a2Dv5b1G00519500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00519600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00519700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00519800.1	Bradi5g10840	Zinc finger
TraesCLr22a2Dv5b1G00519900.1	Bradi5g24820	Probable translation factor pelota ACTIVITY:, endoribonuclease BINDING: protein
TraesCLr22a2Dv5b1G00520000.1	Bradi2g62630	UbiE/COQ5 methyltransferase
TraesCLr22a2Dv5b1G00520100.1	Bradi1g58500	Protein of unknown function DUF1365
TraesCLr22a2Dv5b1G00520200.1	Bradi1g58530	Galactose oxidase/kelch, beta-propeller, BINDING: protein
TraesCLr22a2Dv5b1G00520300.1	Bradi5g24800	Amino acid transporter
TraesCLr22a2Dv5b1G00520400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00520500.1	Bradi1g28640	Cellular retinaldehyde-binding
TraesCLr22a2Dv5b1G00520600.1	Bradi5g24530	Proteinase inhibitor
TraesCLr22a2Dv5b1G00520700.1	Bradi5g24540	Transcription factor
TraesCLr22a2Dv5b1G00520800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00520900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00521000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00521100.1	Bradi5g24570	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00521200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00521300.1	Bradi5g25770	Transcriptional factor
TraesCLr22a2Dv5b1G00521400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00521500.1	Bradi5g25770	Transcriptional factor
TraesCLr22a2Dv5b1G00521600.1	Bradi5g24580	Nucleolar, Nop52
TraesCLr22a2Dv5b1G00521700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00521800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00521900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00522000.1	Bradi5g24590	ATPase
TraesCLr22a2Dv5b1G00522100.1	Bradi3g10810	C2 calcium-dependent membrane targeting
TraesCLr22a2Dv5b1G00522200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00522300.1	Bradi5g24590	ATPase
TraesCLr22a2Dv5b1G00522400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00522500.1	Bradi5g24610	Glycoside hydrolase-type carbohydrate-binding
TraesCLr22a2Dv5b1G00522600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00522700.1	Bradi5g25770	Transcriptional factor
TraesCLr22a2Dv5b1G00522800.1	Bradi5g25770	Transcriptional factor
TraesCLr22a2Dv5b1G00522900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00523000.1	Bradi5g24610	Glycoside hydrolase-type carbohydrate-binding
TraesCLr22a2Dv5b1G00523100.1	Bradi5g24630	Carbamoyl phosphate synthetase
TraesCLr22a2Dv5b1G00523200.1	Bradi5g24640	Polyprenyl synthetase
TraesCLr22a2Dv5b1G00523300.1	No homolog	No homology

TraesCLr22a2Dv5b1G00523400.1	Bradi5g24650	Haem peroxidase, plant/fungal/bacterial, ACTIVITY: electron carrier, peroxidase
TraesCLr22a2Dv5b1G00523500.1	Bradi5g24660	Rossmann-like alpha/beta/alpha sandwich fold
TraesCLr22a2Dv5b1G00523600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00523700.1	Bradi5g24670	Transcriptional factor
TraesCLr22a2Dv5b1G00523800.1	Bradi3g43600	Cytochrome P450
TraesCLr22a2Dv5b1G00523900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00524000.1	Bradi5g24690	ATPase
TraesCLr22a2Dv5b1G00524100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00524200.1	Bradi5g24720	Pathogenesis-related transcriptional factor
TraesCLr22a2Dv5b1G00524300.1	Bradi5g24730	ACTIVITY: structural molecule, protein kinase, BINDING: cytoskeletal protein
TraesCLr22a2Dv5b1G00524400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00524500.1	Bradi1g50560	BINDING: protein
TraesCLr22a2Dv5b1G00524600.1	Bradi3g07390	Ankyrin ACTIVITY: ion channel, catalytic, BINDING: cytoskeletal protein
TraesCLr22a2Dv5b1G00524700.1	Bradi2g11620	Cytochrome P450
TraesCLr22a2Dv5b1G00524800.1	Bradi2g21320	UDP-glucuronosyl/UDP-glucosyltransferase
TraesCLr22a2Dv5b1G00524900.1	Bradi2g21320	UDP-glucuronosyl/UDP-glucosyltransferase
TraesCLr22a2Dv5b1G00525000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00525100.1	Bradi5g24760	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00525200.1	Bradi5g01170	Leucine-rich repeat
TraesCLr22a2Dv5b1G00525300.1	Bradi5g24760	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00525400.1	Bradi2g49090	UDP-glucuronosyl/UDP-glucosyltransferase
TraesCLr22a2Dv5b1G00525500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00525600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00525700.1	Bradi2g40790	Glutathione S-transferase/chloride channel
TraesCLr22a2Dv5b1G00525800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00525900.1	Bradi5g24480	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00526000.1	Bradi5g24470	Translation initiation factor, transcription coactivator
TraesCLr22a2Dv5b1G00526100.1	Bradi5g24450	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00526200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00526300.1	Bradi2g44120	Cytochrome P450
TraesCLr22a2Dv5b1G00526400.1	Bradi3g46790	Major facilitator superfamily
TraesCLr22a2Dv5b1G00526500.1	Bradi1g70860	Basic helix-loop-helix dimerisation region
TraesCLr22a2Dv5b1G00526600.1	Bradi1g05890	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00526700.1	Bradi5g21390	Terpenoid cylases/protein prenyltransferase
TraesCLr22a2Dv5b1G00526800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00526900.1	Bradi4g12110	Protein of unknown function DUF6
TraesCLr22a2Dv5b1G00527000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00527100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00527200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00527300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00527400.1	Bradi1g53680	Zinc/iron permease
TraesCLr22a2Dv5b1G00527500.1	Bradi3g22400	Cellular retinaldehyde-binding/triple function

TraesCLr22a2Dv5b1G00527600.1	Bradi5g24360	Pathogenesis-related transcriptional factor
TraesCLr22a2Dv5b1G00527700.1	Bradi5g24370	Ferredoxin reductase-type FAD-binding domain
TraesCLr22a2Dv5b1G00527800.1	Bradi3g36740	Transcription termination factor
TraesCLr22a2Dv5b1G00527900.1	Bradi4g42520	Major facilitator superfamily
TraesCLr22a2Dv5b1G00528000.1	Bradi5g24380	Aux/IAA-ARF-dimerisation ACTIVITY: protein, dimerization
TraesCLr22a2Dv5b1G00528100.1	Bradi1g57780	Concanavalin A-like lectin/glucanase
TraesCLr22a2Dv5b1G00528200.1	Bradi1g60440	Sodium/calcium exchanger membrane region
TraesCLr22a2Dv5b1G00528300.1	Bradi5g24380	Aux/IAA-ARF-dimerisation ACTIVITY: protein, dimerization
TraesCLr22a2Dv5b1G00528400.1	Bradi5g24380	Aux/IAA-ARF-dimerisation ACTIVITY: protein, dimerization
TraesCLr22a2Dv5b1G00528500.1	Bradi1g55150	Transcription coactivator
TraesCLr22a2Dv5b1G00528600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00528700.1	Bradi1g66540	Heat shock protein Hsp70
TraesCLr22a2Dv5b1G00528800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00528900.1	Bradi4g26860	Staphylococcal nuclease (SNase-like)
TraesCLr22a2Dv5b1G00529000.1	Bradi5g24420	Peptidyl-tRNA hydrolase
TraesCLr22a2Dv5b1G00529100.1	Bradi5g24430	Phospholipase C/P1 nuclease
TraesCLr22a2Dv5b1G00529200.1	Bradi5g24140	Phospholipase C/P1 nuclease
TraesCLr22a2Dv5b1G00529300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00529400.1	Bradi3g43600	Cytochrome P450
TraesCLr22a2Dv5b1G00529500.1	Bradi3g18640	Kelch related ACTIVITY: transcription
TraesCLr22a2Dv5b1G00529600.1	Bradi5g24170	Sulphate transporter
TraesCLr22a2Dv5b1G00529700.1	Bradi5g24180	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00529800.1	Bradi3g43600	Cytochrome P450
TraesCLr22a2Dv5b1G00529900.1	Bradi4g21930	Transferase
TraesCLr22a2Dv5b1G00530000.1	Bradi5g24180	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00530100.1	Bradi5g24180	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00530200.1	Bradi5g24310	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00530300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00530400.1	Bradi5g24190	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00530500.1	Bradi5g24200	Haem peroxidase
TraesCLr22a2Dv5b1G00530600.1	Bradi3g24410	Glycoside hydrolase
TraesCLr22a2Dv5b1G00530700.1	Bradi4g10630	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00530800.1	Bradi1g62010	Alpha-N-acetylglucosaminidase
TraesCLr22a2Dv5b1G00530900.1	Bradi5g24220	D-isomer specific 2-hydroxyacid dehydrogenase
TraesCLr22a2Dv5b1G00531000.1	Bradi4g10630	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00531100.1	Bradi4g10630	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00531200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00531300.1	Bradi4g44590	NB-ARC
TraesCLr22a2Dv5b1G00531400.1	Bradi5g24180	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00531500.1	Bradi5g24200	Haem peroxidase
TraesCLr22a2Dv5b1G00531600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00531700.1	Bradi4g40420	Serine/Thr protein kinase

TraesCLr22a2Dv5b1G00531800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00531900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00532000.1	Bradi5g24220	D-isomer specific 2-hydroxyacid dehydrogenase
TraesCLr22a2Dv5b1G00532100.1	Bradi5g26520	Peptidase
TraesCLr22a2Dv5b1G00532200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00532300.1	Bradi5g24220	D-isomer specific 2-hydroxyacid dehydrogenase
TraesCLr22a2Dv5b1G00532400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00532600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00532500.1	Bradi4g14100	WD40/YVTN repeat-like ACTIVITY: apoptotic, protease activator
TraesCLr22a2Dv5b1G00532700.1	Bradi5g24290	Glycosyl transferase
TraesCLr22a2Dv5b1G00532800.1	Bradi5g22930	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00532900.1	Bradi1g05670	Ribosomal protein S13-like
TraesCLr22a2Dv5b1G00533000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00533100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00533200.1	Bradi1g49600	Nucleotide-binding, alpha-beta plait BINDING:, protein, nucleotide, RNA, DNA
TraesCLr22a2Dv5b1G00533300.1	Bradi5g23700	IQ calmodulin-binding region
TraesCLr22a2Dv5b1G00533400.1	Bradi5g23710	Pentatricopeptide repeat
TraesCLr22a2Dv5b1G00533500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00533600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00533700.1	Bradi5g23720	Plant methyltransferase dimerisation
TraesCLr22a2Dv5b1G00533800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00533900.1	Bradi1g30610	Valyl/Leucyl/Isoleucyl-tRNA synthetase
TraesCLr22a2Dv5b1G00534000.1	Bradi4g07820	Protein of unknown function DUF1649
TraesCLr22a2Dv5b1G00534100.1	Bradi4g07810	Bifunctional dihydrofolate reductase/thymidylate, synthase
TraesCLr22a2Dv5b1G00534200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00534300.1	Bradi5g23760	Ubiquitin-associated/translation elongation
TraesCLr22a2Dv5b1G00534400.1	Bradi2g11840	Pectinesterase inhibitor
TraesCLr22a2Dv5b1G00534500.1	Bradi5g23770	Plant lipid transfer protein
TraesCLr22a2Dv5b1G00534600.1	Bradi5g23790	Alpha/beta hydrolase
TraesCLr22a2Dv5b1G00534700.1	Bradi3g07880	Transcription elongation factor
TraesCLr22a2Dv5b1G00534800.1	Bradi5g14010	Phosphatidylethanolamine-binding, conserved, site BINDING: lipid, ATP
TraesCLr22a2Dv5b1G00534900.1	Bradi4g12950	EGF-like region
TraesCLr22a2Dv5b1G00535000.1	Bradi5g23810	Cystathionine beta-synthase
TraesCLr22a2Dv5b1G00535100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00535200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00535300.1	Bradi5g23880	C2 calcium-dependent membrane targeting
TraesCLr22a2Dv5b1G00535400.1	Bradi5g23890	Tetratricopeptide-like helical ACTIVITY:catalytic
TraesCLr22a2Dv5b1G00535500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00535600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00535700.1	Bradi3g15340	PAK-box/P21-Rho-binding
TraesCLr22a2Dv5b1G00535800.1	Bradi5g23890	Tetratricopeptide-like helical ACTIVITY:catalytic

TraesCLr22a2Dv5b1G00536000.1	Bradi3g14740	Pyridoxal phosphate-dependent transferase
TraesCLr22a2Dv5b1G00535900.1	Bradi3g14740	Pyridoxal phosphate-dependent transferase
TraesCLr22a2Dv5b1G00536100.1	Bradi3g14750	Pyridoxal phosphate-dependent transferase
TraesCLr22a2Dv5b1G00536200.1	Bradi3g58810	Cyclin-like F-box
TraesCLr22a2Dv5b1G00536300.1	Bradi5g23910	Protein of unknown function DUF6
TraesCLr22a2Dv5b1G00536400.1	Bradi5g16660	Short-chain dehydrogenase/reductase
TraesCLr22a2Dv5b1G00536500.1	Bradi5g23920	Peptidase C19
TraesCLr22a2Dv5b1G00536600.1	Bradi1g51270	Leucine zipper
TraesCLr22a2Dv5b1G00536700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00536800.1	Bradi5g27690	Haem peroxidase, plant/fungal/bacterial, ACTIVITY: electron carrier, peroxidase
TraesCLr22a2Dv5b1G00536900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00537000.1	Bradi5g23940	Carbohydrate kinase
TraesCLr22a2Dv5b1G00537100.1	Bradi1g47620	Leucine-rich repeat
TraesCLr22a2Dv5b1G00537200.1	Bradi1g47610	Leucine-rich repeat
TraesCLr22a2Dv5b1G00537300.1	Bradi5g19450	ACTIVITY: GTPase BINDING: protein, GTP
TraesCLr22a2Dv5b1G00537400.1	Bradi5g23970	Zinc finger
TraesCLr22a2Dv5b1G00537500.1	Bradi1g49040	ATPase
TraesCLr22a2Dv5b1G00537600.1	Bradi5g23980	Unknown function
TraesCLr22a2Dv5b1G00537700.1	Bradi5g23980	Unknown function
TraesCLr22a2Dv5b1G00537800.1	Bradi5g10660	Zinc finger
TraesCLr22a2Dv5b1G00537900.1	Bradi3g36980	Mitochondrial glycoprotein
TraesCLr22a2Dv5b1G00538000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00538100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00538200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00538300.1	Bradi5g24120	Zinc finger
TraesCLr22a2Dv5b1G00538400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00538500.1	Bradi5g24110	Pathogenesis-related transcriptional factor
TraesCLr22a2Dv5b1G00538600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00538700.1	Bradi5g24100	Pathogenesis-related transcriptional factor
TraesCLr22a2Dv5b1G00538800.1	Bradi5g24090	Protein of unknown function DUF869
TraesCLr22a2Dv5b1G00538900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00539000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00539100.1	Bradi5g24060	Serine/Thr protein kinase-related
TraesCLr22a2Dv5b1G00539200.1	Bradi5g24060	Serine/Thr protein kinase-related
TraesCLr22a2Dv5b1G00539300.1	Bradi5g22880	Leucine-rich repeat
TraesCLr22a2Dv5b1G00539400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00539500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00539600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00539700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00539800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00539900.1	Bradi1g51270	Leucine zipper
TraesCLr22a2Dv5b1G00540000.1	Bradi5g24060	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00540100.1	Bradi5g24040	Protein of unknown function DUF579

TraesCLr22a2Dv5b1G00540200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00540300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00540500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00540600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00540700.1	Bradi5g18990	Dihydrodipicolinate synthase subfamily
TraesCLr22a2Dv5b1G00540800.1	Bradi3g50860	ACTIVITY: protein, homodimerization, protein heterodimerization
TraesCLr22a2Dv5b1G00540900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00541000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00541100.1	Bradi5g24020	Proteinase inhibitor
TraesCLr22a2Dv5b1G00541200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00541300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00541400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00541500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00541600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00541700.1	Bradi3g16300	Alpha/beta hydrolase fold-1
TraesCLr22a2Dv5b1G00541800.1	Bradi5g23420	Protein of unknown function DUF617
TraesCLr22a2Dv5b1G00541900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00542000.1	Bradi5g23420	Protein of unknown function DUF617
TraesCLr22a2Dv5b1G00542100.1	Bradi5g23450	Protein kinase-like ACTIVITY:, 2-octaprenylphenol hydroxylase
TraesCLr22a2Dv5b1G00542200.1	Bradi5g23460	Shikimate kinase
TraesCLr22a2Dv5b1G00542300.1	Bradi1g54260	F-box associated type 1
TraesCLr22a2Dv5b1G00542400.1	Bradi5g23470	Glycoside hydrolase
TraesCLr22a2Dv5b1G00542500.1	Bradi5g23500	Programmed cell death protein 2
TraesCLr22a2Dv5b1G00542600.1	Bradi5g23510	Calcium-binding EF-hand ACTIVITY: transcription, regulator, receptor
TraesCLr22a2Dv5b1G00542700.1	Bradi2g05780	DNA-directed RNA polymerase
TraesCLr22a2Dv5b1G00542800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00542900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00543000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00543100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00543200.1	Bradi1g53170	Cyclin-like F-box
TraesCLr22a2Dv5b1G00543300.1	Bradi3g00340	E3 ubiquitin ligase
TraesCLr22a2Dv5b1G00543400.1	Bradi3g00340	E3 ubiquitin ligase
TraesCLr22a2Dv5b1G00543500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00543600.1	Bradi5g23550	Pectin lyase fold/virulence factor
TraesCLr22a2Dv5b1G00543700.1	Bradi2g53320	Cation/H+ exchanger
TraesCLr22a2Dv5b1G00543800.1	Bradi5g23560	Thioredoxin-like fold ACTIVITY: electron, carrier, protein disulfide oxidoreductase
TraesCLr22a2Dv5b1G00543900.1	Bradi5g23570	Nucleotide-binding, alpha-beta plait, ACTIVITY: oxidoreductase
TraesCLr22a2Dv5b1G00544000.1	Bradi5g23600	ABC transporter
TraesCLr22a2Dv5b1G00544100.1	Bradi3g43600	Cytochrome P450
TraesCLr22a2Dv5b1G00544200.1	Bradi5g23610	Calponin-like actin-binding
TraesCLr22a2Dv5b1G00544300.1	Bradi4g34060	Protein of unknown function DUF599

TraesCLr22a2Dv5b1G00544400.1	Bradi5g13260	Glycoside hydrolase
TraesCLr22a2Dv5b1G00544500.1	Bradi5g20060	Bet v I allergen
TraesCLr22a2Dv5b1G00544600.1	Bradi5g23650	Winged helix repressor DNA-binding ACTIVITY:, nuclease, ubiquitin-protein ligase
TraesCLr22a2Dv5b1G00544700.1	Bradi4g23550	Peptidase S8 and S53
TraesCLr22a2Dv5b1G00544800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00544900.1	Bradi4g03440	Cytochrome P450
TraesCLr22a2Dv5b1G00545000.1	Bradi1g02820	Proteinase inhibitor I29
TraesCLr22a2Dv5b1G00545100.1	Bradi5g23660	ATPase, V1/A1 complex, subunit D ACTIVITY:, proton-transporting ATPase
TraesCLr22a2Dv5b1G00545200.1	Bradi5g23670	Glutamyl-tRNA(Gln) amidotransferase A subunit, ACTIVITY: glutaminyl-tRNA synthase
TraesCLr22a2Dv5b1G00545300.1	Bradi5g23680	3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase
TraesCLr22a2Dv5b1G00545400.1	Bradi1g75060	Protein of unknown function DUF1677
TraesCLr22a2Dv5b1G00545500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00545600.1	Bradi4g42470	Cyclin-like F-box
TraesCLr22a2Dv5b1G00545700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00545800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00545900.1	Bradi1g31760	Pentatricopeptide repeat
TraesCLr22a2Dv5b1G00546000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00546100.1	Bradi5g09910	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00546200.1	Bradi1g45220	Transcription factor TCP subgroup
TraesCLr22a2Dv5b1G00546300.1	Bradi1g72700	Protein of unknown function DUF567
TraesCLr22a2Dv5b1G00546400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00546500.1	No homolog	No homology
TraesCLr22a2Dv5b1G00546600.1	Bradi5g23400	Transferase ACTIVITY: transferase, transferring, acyl groups other than amino-acyl groups
TraesCLr22a2Dv5b1G00546700.1	Bradi5g23400	Transferase ACTIVITY: transferase, transferring, acyl groups other than amino-acyl groups
TraesCLr22a2Dv5b1G00546800.1	No homolog	No homology
TraesCLr22a2Dv5b1G00546900.1	Bradi5g23400	Transferase ACTIVITY: transferase, transferring, acyl groups other than amino-acyl groups
TraesCLr22a2Dv5b1G00547000.1	Bradi1g11680	Lipase/lipooxygenase
TraesCLr22a2Dv5b1G00547100.1	Bradi1g29200	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00547200.1	Bradi5g23400	Transferase ACTIVITY: transferase, transferring, acyl groups other than amino-acyl groups
TraesCLr22a2Dv5b1G00547300.1	Bradi4g08150	DNA/RNA helicase
TraesCLr22a2Dv5b1G00547400.1	Bradi5g23340	Basic-leucine zipper (bZIP) transcription, factor ACTIVITY: protein dimerization
TraesCLr22a2Dv5b1G00547500.1	Bradi5g23330	Regulator of chromosome condensation
TraesCLr22a2Dv5b1G00547600.1	No homolog	No homology
TraesCLr22a2Dv5b1G00547700.1	Bradi5g23320	unknown function
TraesCLr22a2Dv5b1G00547800.1	Bradi4g30430	Zinc finger
TraesCLr22a2Dv5b1G00547900.1	Bradi5g23310	BTB/POZ fold ACTIVITY: signal transducer, BINDING: protein
TraesCLr22a2Dv5b1G00548000.1	Bradi3g50880	N-6 adenine-specific DNA methylase
TraesCLr22a2Dv5b1G00548100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00548200.1	Bradi5g23280	Phospholipase C/P1 nuclease

TraesCLr22a2Dv5b1G00548300.1	Bradi5g23270	Pyridoxal phosphate-dependent transferase
TraesCLr22a2Dv5b1G00548400.1	Bradi5g23270	Pyridoxal phosphate-dependent transferase
TraesCLr22a2Dv5b1G00548600.1	Bradi3g60340	NB-ARC
TraesCLr22a2Dv5b1G00548700.1	Bradi5g23250	Glycosyl transferase
TraesCLr22a2Dv5b1G00548800.1	Bradi5g23260	RNA polymerase II transcription factor
TraesCLr22a2Dv5b1G00548900.1	No homolog	No homology
TraesCLr22a2Dv5b1G00549000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00549100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00549200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00549300.1	Bradi2g32640	Ribosomal protein L5 BINDING: protein, structural constituent of ribosome
TraesCLr22a2Dv5b1G00549400.1	Bradi2g39720	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00549500.1	Bradi2g01170	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00549600.1	Bradi4g17230	Chalcone and stilbene synthases
TraesCLr22a2Dv5b1G00549800.1	Bradi2g01320	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00549700.1	Bradi2g39720	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00549900.1	Bradi3g40050	Short-chain dehydrogenase/reductase
TraesCLr22a2Dv5b1G00550000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00550100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00550200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00550300.1	Bradi4g03140	Peptidase, trypsin-like Ser and cysteine
TraesCLr22a2Dv5b1G00550500.1	Bradi5g23240	Unknown function
TraesCLr22a2Dv5b1G00550400.1	Bradi5g23230	Metallophosphoesterase
TraesCLr22a2Dv5b1G00550600.1	Bradi5g23220	Like-Sm ribonucleoprotein, eukaryotic and, archaea-type, core ACTIVITY: protein
TraesCLr22a2Dv5b1G00550700.1	Bradi5g23190	Aldehyde dehydrogenase
TraesCLr22a2Dv5b1G00550800.1	Bradi5g23200	Pentatricopeptide repeat
TraesCLr22a2Dv5b1G00550900.1	Bradi5g23210	GCN5-related N-acetyltransferase
TraesCLr22a2Dv5b1G00551000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00551100.1	Bradi3g22560	Cytochrome P450
TraesCLr22a2Dv5b1G00551300.1	Bradi2g38440	Pentatricopeptide repeat
TraesCLr22a2Dv5b1G00551200.1	Bradi2g38440	Pentatricopeptide repeat
TraesCLr22a2Dv5b1G00551400.1	Bradi4g38970	Glycosyl transferase
TraesCLr22a2Dv5b1G00551500.1	Bradi4g38980	Aminoacyl-tRNA synthetase
TraesCLr22a2Dv5b1G00551600.1	Bradi1g17530	Cytochrome c oxidase
TraesCLr22a2Dv5b1G00551700.1	Bradi5g23130	Diacylglycerol kinase
TraesCLr22a2Dv5b1G00551800.1	Bradi5g23130	Diacylglycerol kinase
TraesCLr22a2Dv5b1G00551900.1	Bradi5g23130	Diacylglycerol kinase
TraesCLr22a2Dv5b1G00552000.1	No homolog	No homology
TraesCLr22a2Dv5b1G00552100.1	Bradi5g23130	Diacylglycerol kinase
TraesCLr22a2Dv5b1G00552200.1	No homolog	No homology
TraesCLr22a2Dv5b1G00552300.1	No homolog	No homology
TraesCLr22a2Dv5b1G00552400.1	No homolog	No homology
TraesCLr22a2Dv5b1G00552500.1	Bradi5g23130	Diacylglycerol kinase

TraesCLr22a2Dv5b1G00552700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00552600.1	Bradi5g23110	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00552800.1	Bradi5g23130	Diacylglycerol kinase
TraesCLr22a2Dv5b1G00552900.1	Bradi5g23120	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00553000.1	Bradi2g59060	Exo70 exocyst complex
TraesCLr22a2Dv5b1G00553100.1	Bradi1g40150	Reticulon BINDING: protein
TraesCLr22a2Dv5b1G00553200.1	Bradi5g22840	Leucine-rich repeat
TraesCLr22a2Dv5b1G00553300.1	Bradi1g78530	Peptidase M10A and M12B
TraesCLr22a2Dv5b1G00553400.1	Bradi2g19520	Lecithin:cholesterol acyltransferase
TraesCLr22a2Dv5b1G00553500.1	Bradi5g23110	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00553600.1	Bradi4g42540	Cytochrome P450
TraesCLr22a2Dv5b1G00553700.1	No homolog	No homology
TraesCLr22a2Dv5b1G00553800.1	Bradi1g20040	Alkyl hydroperoxide reductase/ Thiol specific
TraesCLr22a2Dv5b1G00553900.1	Bradi1g20050	Six-bladed beta-propeller, TolB-like
TraesCLr22a2Dv5b1G00554000.1	Bradi5g00980	Cytochrome P450
TraesCLr22a2Dv5b1G00554100.1	No homolog	No homology
TraesCLr22a2Dv5b1G00554200.1	Bradi1g36640	Haem oxygenase
TraesCLr22a2Dv5b1G00554300.1	Bradi5g23110	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00554400.1	Bradi3g00300	Basic-leucine zipper (bZIP) transcription factor
TraesCLr22a2Dv5b1G00554500.1	Bradi5g23080	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00554600.1	Bradi1g15300	Protein of unknown function DUF952

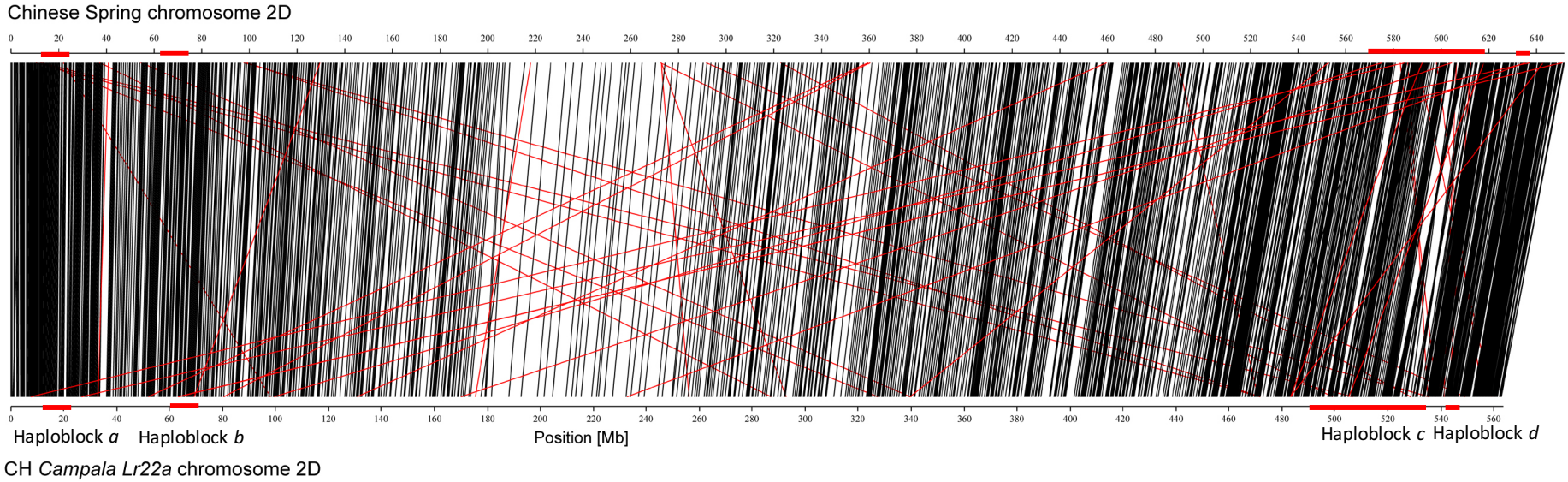
Supplementary table S3.2. Unique genes identified in Chinese Spring and CH Campala *Lr22a* along with the best *Brachypodium distachyon* gene hit and the functional annotation based on *Brachypodium distachyon* genes.

Unique gene Chinese Spring	Best Brachipodium hit	Function
TraesCS2D01G501600	Bradi3g56100	DNA/RNA helicase
TraesCS2D01G527000	Bradi1g15970	Per1-like family protein
TraesCS2D01G511700	Bradi1g22500	Leucine-rich repeat
TraesCS2D01G510000	Bradi1g29410	Binding protein
TraesCS2D01G510300	Bradi1g29410	Binding protein
TraesCS2D01G468300	Bradi1g34070	Anthranilate synthase component I and chorismate binding protein
TraesCS2D01G501300	Bradi1g40150	Reticulon binding protein
TraesCS2D01G008400	Bradi1g65760	Lateral organ boundaries (LOB)
TraesCS2D01G509900	Bradi1g76920	Serine/Thr protein kinase
TraesCS2D01G584500	Bradi2g13090	ATPase, F1/V1/A1 complex, alpha/beta subunit, nucleotide-binding
TraesCS2D01G569100	Bradi2g19290	Serine/Thr protein kinase
TraesCS2D01G522000	Bradi2g37820	DNA repair and recombination
TraesCS2D01G499900	Bradi2g41790	Ionotropic glutamate-like receptor,
TraesCS2D01G052900	Bradi2g51570	Serine/Thr protein kinase
TraesCS2D01G517000	Bradi2g60970	Basic helix-loop-helix dimerisation region
TraesCS2D01G520400	Bradi2g61870:	Multi antimicrobial extrusion protein MatE
TraesCS2D01G045200	Bradi3g36730	Zinc finger
TraesCS2D01G528600	Bradi3g57890	Mitochondrial substrate carrier
TraesCS2D01G510100	Bradi4g06970	Leucine-rich repeat
TraesCS2D01G574800	Bradi4g12750	Translation elongation factor EF1A/initiation, factor IF2gamma
TraesCS2D01G499400	Bradi5g25770	Transcriptional factor B3
TraesCS2D01G552300	Bradi5g25890	Plastocyanin-like, domain-containing protein
TraesCS2D01G149600	No Hit	
TraesCS2D01G513400	No Hit	
TraesCS2D01G521900	No Hit	
TraesCS2D01G573100	No Hit	

Unique gene CH Campala *Lr22a*	Best Brachipodium hit	Function
TraesCLr22a2Dv5b1G00508900.1	Bradi3g02240	Leucine-rich repeat
TraesCLr22a2Dv5b1G00467400.1	Bradi1g07700	Peptidase S8 and S53
TraesCLr22a2Dv5b1G00554800.1	Bradi1g14350	Multi antimicrobial extrusion protein MatE
TraesCLr22a2Dv5b1G00383900.1	Bradi1g43160	Short-chain dehydrogenase/reductase
TraesCLr22a2Dv5b1G00399500.1	Bradi1g49410	Peptidase aspartic
TraesCLr22a2Dv5b1G00399400.1	Bradi1g49460	Pentatricopeptide repeat
TraesCLr22a2Dv5b1G00524500.1	Bradi1g50560	BINDING: protein
TraesCLr22a2Dv5b1G00399700.1	Bradi1g50700	Protein of unknown function DUF594
TraesCLr22a2Dv5b1G00507800.1	Bradi1g56830	Alpha/beta hydrolase
TraesCLr22a2Dv5b1G00586800.1	Bradi2g14120	zinc finger
TraesCLr22a2Dv5b1G00553400.1	Bradi2g19520	Lecithin:cholesterol acyltransferase
TraesCLr22a2Dv5b1G00549300.1	Bradi2g32640	Ribosomal protein L5 BINDING: protein
TraesCLr22a2Dv5b1G00606700.1	Bradi2g42410	DNA/RNA helicase, ATP-dependent, DEAH-box type, conserved site
TraesCLr22a2Dv5b1G00065900.1	Bradi2g43440	LPPG:FO 2-phospho-L-lactate transferase
TraesCLr22a2Dv5b1G00514400.1	Bradi2g46250	Glycosyl transferase
TraesCLr22a2Dv5b1G00063400.1	Bradi2g51150	Uncharacterised protein family UPF0089
TraesCLr22a2Dv5b1G00582900.1	Bradi2g60350	Guanine nucleotide binding protein
TraesCLr22a2Dv5b1G00565600.1	Bradi3g03400	Cyclin-like F-box
TraesCLr22a2Dv5b1G00045200.1	Bradi3g12470	Peptidase S8 and S53
TraesCLr22a2Dv5b1G00606900.1	Bradi3g16810	Protein kinase
TraesCLr22a2Dv5b1G00511100.1	Bradi3g35170	Harpin-induced 1
TraesCLr22a2Dv5b1G00256000.1	Bradi3g42710	Cytochrome c
TraesCLr22a2Dv5b1G00255900.1	Bradi3g42730	Protein-Tyr phosphatase
TraesCLr22a2Dv5b1G00256100.1	Bradi3g42740	Beta tubulin
TraesCLr22a2Dv5b1G00511400.1	Bradi3g47790	Phosphopantetheine attachment site
TraesCLr22a2Dv5b1G00518800.1	Bradi4g24650	ABA/WDS induced protein
TraesCLr22a2Dv5b1G00048600.1	Bradi4g26620	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00383800.1	Bradi4g37860	Male sterility
TraesCLr22a2Dv5b1G00038600.1	Bradi4g39940	Peptidase S10
TraesCLr22a2Dv5b1G00531300.1	Bradi4g44590	NB-ARC
TraesCLr22a2Dv5b1G00002400.1	Bradi5g00530	zinc finger
TraesCLr22a2Dv5b1G00038500.1	Bradi5g02510	Protein of unknown function DUF1486
TraesCLr22a2Dv5b1G00043100.1	Bradi5g02910	Pumilio RNA-binding region
TraesCLr22a2Dv5b1G00043000.1	Bradi5g02920	Aminoacyl-tRNA synthetase
TraesCLr22a2Dv5b1G00360600.1	Bradi5g13860	Glucose-6-phosphate dehydrogenase

TraesCLr22a2Dv5b1G00002500.1	Bradi3g03110	Serine/Thr protein kinase
TraesCLr22a2Dv5b1G00050100.1	No Hit	
TraesCLr22a2Dv5b1G00283600.1	No Hit	
TraesCLr22a2Dv5b1G00502100.1	No Hit	
TraesCLr22a2Dv5b1G00507700.1	No Hit	
TraesCLr22a2Dv5b1G00546400.1	No Hit	
TraesCLr22a2Dv5b1G00552300.1	No Hit	
TraesCLr22a2Dv5b1G00586900.1	No Hit	

Supplementary figure S3.3. Gene collinearity between Chinese Spring and 'CH Campala *Lr22a*'. Collinear genes are connected with black lines and non-collinear genes are connected with red lines. For better visibility, only every fifth gene is displayed. The purpose is to illustrate that the vast majority of genes are in perfectly collinear order. At the centromeric region (190-290 Mb in Chinese Spring and 150-250 Mb in 'CH Campala *Lr22a*'), the gene density is low but they show good collinearity.



Chapter 4

General Discussion and Outlook

To achieve sustainable agriculture, disease resistance is an important prerequisite. In the past, many genes were cloned using traditional map-based cloning, a process that often lasted for up to a decade in wheat. There are ~430 disease resistance genes that have been identified and that are described in the wheat gene catalogue (McIntosh et al., 1995). However, only 25 of these genes have been cloned. Map-based cloning is a time-consuming and labor-intensive approach and is not an ideal solution if we want to clone a large proportion of the described genes. But before discussing how to best clone genes, it is important to understand why we need to clone them.

There are various reasons why gene cloning is important. The first is to determine the function of disease resistance genes at the molecular level for strategic deployment of resistance genes for sustainable agriculture. A second reason is that gene cloning enables direct gene transfer between wheat cultivars and cereal species via transgenesis. Gene cassettes are an ideal tool to transfer genes between sexually incompatible species and also to combine functionally variable alleles of a given resistance genes (Wulff & Moscou, 2014). Third, the cloning of genes opens possibilities to design perfectly diagnostic molecular markers on the functional polymorphisms that distinguish resistant from susceptible alleles, and fourth in the future it will allow the use of genome editing to introduce functional polymorphisms in susceptible wheat cultivars or other cereal species. Cloning of resistance genes such as *Lr22a* offers the exciting possibility to elucidate the molecular mechanisms of broad-spectrum disease resistance. One of the advantages of working with wheat or cereals is that we have access to breeding records from many different countries that often date back for decades. These records provide exact information on genes that showed broad-specificities against many pathogen strains. Comparable information is often lacking in model organisms such as *Arabidopsis*. Our knowledge on the molecular basis of broad-spectrum disease resistance is still very limited. Molecular and biochemical studies will deliver answers to important questions such as: why are some resistance

genes broad-spectrum and other not? How are resistance proteins activated by effector recognition? What are the induced host components that affect immunity? And what are the targets of effectors? This information is lacking in cereal and our current knowledge is from the molecular and biochemical work done in model organism, *Arabidopsis*.

In the following paragraphs, novel gene cloning approaches and how they are used for gene cloning in wheat and barley will be discussed.

4.1 Novel rapid gene cloning approaches

The advancement in DNA sequencing technologies and in the development of genome complexity reduction accelerated gene cloning in cereals with large genomes. In particular, these approaches aimed to eliminate the need for chromosome walking, which traditionally has been one of the most time-consuming steps in map-based gene cloning projects. Recently, four gene cloning technologies have been developed that facilitate rapid gene cloning in wheat and barley, MutRenSeq, MutChromSeq, TACCA and AgRenSeq.

MutRenSeq is a modified version of 'Resistance gene enrichment Sequencing' (RenSeq) which involves capturing fragments from genomic and cDNA libraries using biotinylated RNA oligonucleotides which are designed to be complementary to NLR-encoding genes (Jupe et al., 2013). MutRenSeq is used to clone NLRs only, as most *R* genes encode for NLRs (Steuernagel et al., 2016). MutRenSeq made use of publically available annotations and RNA-Seq data to design the capture array (Steuernagel et al., 2016). MutRenSeq was used to clone the two stem rust resistance genes *Sr22* and *Sr45* from hexaploid wheat (Steuernagel et al., 2016). MutRenSeq is a three-step process for rapid gene isolation, (i) development of a mutant population from a resistant wild-type parent, identifying loss-of-function mutants and performing NLR capture. For example, for the cloning of *Sr22*, six independent mutants were identified from 1,300 M_2 families. (ii) Sequencing of the wild-type resistant plant and the loss of function mutants using

Illumina short-read sequencing, and (iii) comparing the genes in the wild-type and mutants to identify the mutation that led to the loss of disease resistance. For example, to clone the *Sr22* gene, 23 contigs were identified that were mutated in two mutants, three contigs that were mutated in three mutants and 1 contig of 3,408 bp that contained independent mutations in five of the six mutants. This 3,408 bp contig showed homology with the C-terminus of an *Ae. tauschii* NLR homolog (Steuernagel et al., 2016). Using the 5' sequence of this *Ae. tauschii* NLR homolog, a contig containing an EMS induced mutation in the N-terminus of the sixth mutant was identified. The two contigs were physically joined using a PCR of genomic and cDNA to obtain the full-length sequence of the *Sr22* gene and the mutations were confirmed using by Sanger sequencing. To further verify the *Sr22* cloning, a PCR marker was designed on the sequence of the *Sr22* coding sequence that co-segregated with the *Sr22*-mediated resistance phenotype in 2,300 gametes (Steuernagel et al., 2016). As a proof of concept, this approach was first used to clone the well-known *Sr33* gene in a fraction of time compared to the conventional gene isolation techniques (Steuernagel et al., 2016). For *Sr33*, 8,235 genomic contigs (14.5 Mb) enriched in NLR were obtained using Illumina short-read sequencing which resulted in 1,000-fold reduction in genome complexity. MutRenSeq is a fast, cost-effective cloning technique where no fine mapping and generation of the physical region across the target region is required.

AgRenSeq is an innovative advancement of RenSeq that combines association genetics with NLR gene enrichment sequencing (AgRenSeq) to identify the sequences of functional *R* genes in a diversity panel. However, so far it has only been tested in *Ae. tauschii* and it's not clear yet how successful this technique will be in polyploid species. The quality of the diversity panel is an important prerequisite, this can be tested by using the molecular markers to identify and eliminate any redundant accessions. As a proof of concept, AgRenSeq was used to clone the already known stem rust resistance gene *Sr33* and two novel genes *Sr46* and *SrTA1662* in a

diversity panel of *Ae. tauschii* (Arora et al., 2018). It permits the unprecedented immediate identification of functional NLRs in an enrichment-sequenced diversity panel following phenotyping with a pathogen isolate. For AgRenSeq, no mapping population or mutagenesis is required to clone the *R* genes. Also, AgRenSeq directly identifies NLR underlying resistance rather than a genomic region with multiple paralogs which requires candidate gene validation (Sanu Arora et al., 2018). Hence, AgRenSeq can be used to isolate R genes from wild relatives which have varied agronomy and requires only the phenotyping of the enrichment-sequenced diversity panel (S. Arora et al., 2017). However, AgRenseq, cannot be used to clone a particular R gene, it is rather used to pull out R genes distributed in a diversity panel. Also, for cloning of genes from different species, different diversity panel are required, for example for isolating R genes from *T. monococcum, Ae. tauschii* diversity panel cannot be used and one has to generate new diversity panel with diverse *T. monococcum* accessions

MutChromSeq was used to clone a powdery mildew resistance gene, *Pm2*, in wheat (Sanchez-Martin et al., 2016). This technique involves a step of complexity reduction by 'chromosome flow sorting' in which only the target chromosome carrying the gene of interest from the resistant cultivar and EMS-induced loss-of-function mutants is isolated and sequenced. Labelling of the repetitive DNA on chromosomes before flow cytometric chromosome analysis allows purification of the individual chromosomes from wheat and barley. Flow-sorting allows to reduce the size of the genome fraction by a factor of 21 in hexaploid wheat and also eliminates the problem that the presence of homoeologous chromosomes often confounds gene cloning. The sequence of the flow sorted chromosome from wild-type and the mutants is analyzed to identify EMS induced mutations that could be responsible for susceptible phenotype. For example, for *Pm2*, six mutants and the wild-type parent were used. Chromosome 5D was isolated from mutants and the wild-type parent and sequenced on an Illumina platform to 35 x coverage. Sequence analysis revealed two candidate contigs of >1kb in size which were mutated in

all six mutants. One of the contigs was discarded due to high SNPs compared to wild-type, which indicates assembly artefacts whereas the second contig contained a full-length NLR-type resistance gene that was verified as *Pm2* based on a sequence specific PCR marker. Compared to MutRenSeq, for MutChromseq each mutant library is sequenced on a single lane of the Illumina Hi-Seq whereas for MutRenSeq, all mutants are sequenced on a single lane. This makes MutRenSeq more cost-effective than MutChromSeq. On the other hand, MutChromSeq does not make any assumptions about the gene product and can be used to clone any gene. MutChromSeq can be used for species that can be mutagenized such as wheat and barley and where the target gene produces a clear phenotype (Sanchez-Martin et al., 2016).

Targeted chromosome based cloning via long-range assembly (TACCA) is a rapid and cost-effective approach for cloning of genes with partial resistance phenotype. This technique was used to clone partial adult plant resistance gene *Lr22a* (Thind et al., 2017) and has been described in detail in chapter 2.

All these approaches have proved their significance independently by demonstrating the rapid cloning of different disease resistance gene. But what is the best approach to clone a disease resistance gene, given that we have all these tools available now?

4.2 How to clone disease resistance genes in wheat and barley?

Before starting any gene cloning project, it is important to consider the phenotype of the target gene and resources available (Fig. 4.1). The phenotype of the gene determines if EMS mutant screening is likely to succeed or not. In case of a complete resistance phenotype that is already expressed at early seedling stage, susceptible loss-of-function mutants are easy to identify in a greenhouse assay, as the example of the *Pm2* showed. Identification of loss-of-function mutants for *Lr22a* on the other hand was challenging because the *Lr22a* resistance was only

expressed at adult plant stage. In addition, the resistance conferred by *Lr22a* was partial. Screening of gene with complete resistance phenotype expressed at seedling stage is feasible in greenhouses or growth cabinets. Greenhouse assays for adult plant resistance on the other hand are challenging. Hence, EMS populations for APR genes often needs to be scored in the field, which greatly increases the dependence on environmental conditions. Partial resistance phenotypes can make the identification of single susceptible plants difficult. Hence, for disease resistance genes with clear phenotypes, MutRenSeq and MutChromSeq are the methods or choice. If MutRenSeq fails for genes with complete resistance phenotype (when the *R* gene does not encode for NLR), MutChromSeq can be used.

However, for genes with partial resistance phenotype, where mutant identification might become difficult or impossible, the TACCA approach can be used as a valuable addition. TACCA requires a genetic mapping population and it is flexible in terms of the method used for gene validation (Fig. 4.1). Mutant screening is challenging when cloning genes from wild species because of the varied agronomic traits of many wild species such as long generation time, seed shattering and seed dormancy. To overcome this limitation, AgRenSeq can be used instead to identify the sequences of functional *R* genes in a diversity panel of wild species.

Each technique has its own advantages and disadvantages, for example, MutRenSeq and AgRenSeq can only be used to isolate NLR-type resistance genes and not for a novel class of gene family like *Stb6,* which encodes for receptor-like-kinase protein, whereas, TACCA and MutChromSeq can be used to clone any gene where preliminary map information of the gene is available. Another major limitation of MutRenSeq is the misassembly of large NLR genes due to short reads generated by Illumina sequencing. Also for MutChromSeq, where the reference contig is generated through *de novo* assembly of short Illumina reads, the challenge lies in assembling repeated genes that are part of large tandem clusters as they tend to collapse in one contig. In some cases, the gene splits on two contigs due to the separation of exons by large

introns and this complicates the identification of causal loss-of-function mutations. The best strategy to overcome these independent limitations associated with the gene cloning technologies is to use a combination of two or more technologies. For example, assembly

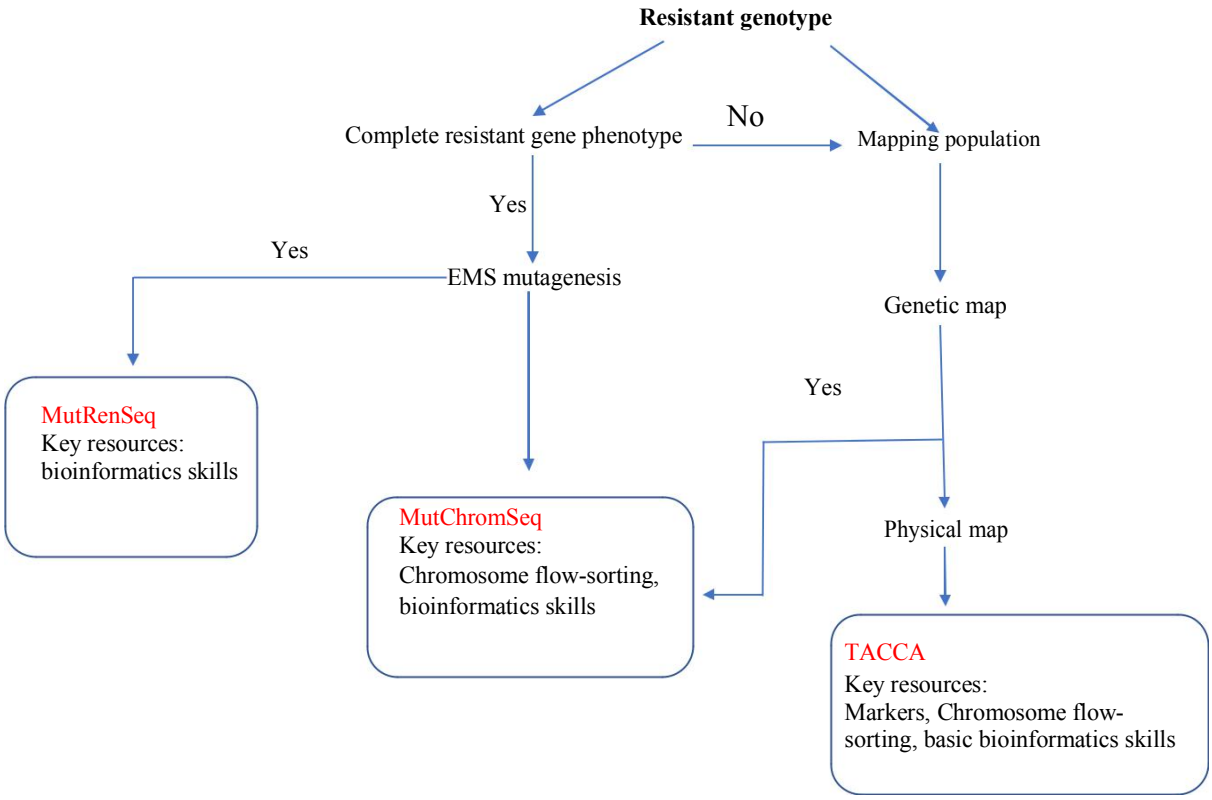


Fig. 4.1 Schematic representation of choice of novel technology for gene cloning in wheat and barley.

problems of MutChromSeq can be overcome by combining MutChromSeq with TACCA to generate a high-quality assembly of the resistant wild-type parent. However, for TACCA cloning, high molecular weight DNA is required, which can take 2-3 months for chromosome flow-sorting whereas for MutChromSeq, amplified chromosomal DNA is used for which tens of thousands of copies of sorted chromosome are required which can be purified in less than one day (Sanchez-Martin et al., 2016).

4.3 *Lr22a* – a quantitative NLR?

A surprising fact about *Lr22a* is its broad-spectrum specificity against all the tested leaf rust isolates. This is because NLRs are typically associated with race-specific resistance. This raises an important and yet unanswered question as to why *Lr22a* shows broad-spectrum resistance? There can be three hypotheses to explain this broad-spectrum resistance. First, this gene has only been used in a limited number of wheat cultivars, particularly in Canadian cultivar AC Minto, 5500HR and 5600HR (Hiebert et al., 2007; McCallum et al., 2016). Therefore, it is possible that this gene has not been sufficiently exposed to leaf rust, which would have resulted in pathogen adaptation.

Second, as *Lr22a* shows homology to the *RPM1* gene in Arabidopsis, which is a classic example of resistance according to the guard model, it is possible that *Lr22a* works analogous to *RPM1* and Lr22a guards the ortholog of RIN4 in wheat. The mechanism can be that the pathogen (*P. tritici*) attacks the wheat plant, the effectors from the pathogen modify the wheat RIN4 (wRIN4) and the perturbations caused in wRIN4 mediate an *Lr22a* specific broad-spectrum disease resistance response in wheat. RIN4 is a regulator of basal defense, therefore even if the pathogen effector evolved to evade *Lr22a* mediated recognition, activation of RIN4 mediated basal defense would still prevent the growth of the pathogen. However, it is also possible that Lr22a does not interact with wRIN4 but has another, novel interactor. This could be tested using a non-targeted yeast-two-hybrid approach in parallel to targeted yeast-two-hybrid experiments specifically testing for Lr22a-wRIN4 interactions.

The third hypothesis is that the broad-spectrum resistance could be due to the specificity-determining residues and the downstream resistance signaling components. Based on the sequence analysis of resistance genes and domain swap experiments between alleles, it was found that the CC and LRR domain determine specificity. The LRR is the major domain

that is required for recognition function whereas the CC domain plays a role in signaling (Kasmia & Nishimurab, 2016) (Ellis et al., 2007). Any change in the residues in the domains such as CC/TIR (Toll/interleukin-1 receptor), NBARC or LRR can have a significant effect on NLR protein structure and function. For example, deletion of the signal anchor sequence of the flax L6 protein destabilizes protein accumulation and renders it non-functional. In wheat, *Pm3* alleles *Pm3a* and *Pm3b* exhibit broad-spectrum resistance compared to allele *Pm3f* and this is due to two amino acids in the ARC2 domain of the NBS. Combined substitution of these two amino acids in the *Pm3f* enhanced HR in *Nicotiana benthamiana* and also enlarged the resistance spectrum of *Pm3f* (Stirnweis et al., 2014). Similarly, there are eleven flax *L* genes, of which ten encode for different flax rust resistance specificities. Sequence analysis showed that variation between these alleles is spread throughout different domains with maximum variation in the LRR-coding region. Comparison of nucleotide sequence of the closely related pair of alleles reveal the residues important for differences in gene-for-gene specificity. For example, L6 and L11 proteins which were identical in the TIR and NBS region differed by 33 amino acids in the LRR region, which indicates that the difference in resistance specificities is caused by the differences in the LRR regions (Ellis et al., 2007). A similar study could be envisaged for *Lr22a* where identification and sequence analysis of resistant and susceptible alleles of *Lr22a* can be used to determine residues which could be responsible for recognizing the leaf rust master effector that cannot be deleted without severely compromising pathogen fitness.

Moreover, *Lr22a* has been reported as a partial resistance gene (Hiebert et al., 2007) whereas we observed a wide range of phenotypes from partial to complete resistance against different Swiss leaf rust isolates. Previously, Thatcher near isogenic lines (NILs) carrying *Lr22a* were screened with leaf rust isolates in Canada, where they showed a partial resistance phenotype against all tested isolates (Hiebert et al., 2007; McCallum et al., 2016).

Surprisingly, the CH Campala NILs carrying *Lr22a* (CH Campala *Lr22a*) showed a complete resistance phenotype, which indicates the genotype-by-genotype interactions with other genes in the background.

Molecular studies will deliver an understanding of the *Lr22a* activation mechanism and provide insights into the broad-spectrum specificity of *Lr22a*. If the guard hypothesis (hypothesis 2) is correct, understanding of the molecular mechanism will result in the identification of a protein that potentially controls basal immunity in wheat or to identify novel interactors or effector molecules. This will also enable us to have insights into how the interactors recognize multiple pathogen effectors leading to broad spectrum resistance. In case there is a direct interaction between the effector and Lr22a, this effector sequence can be used to amplify other closely related effectors from different isolates of *P. triticina.* These effectors can then be used to study interaction with Lr22a using transient expression assay in *Nicotiana benthamiana,* which would allow identification of effectors which can evade *Lr22a* mediated resistance. In both the scenarios, results are expected to guide the way for the activation studies of other plant NLRs and elucidate how NLRs in plant innate immune system work. Moreover, if hypothesis 2 or 3 are correct, *Lr22a* might be used as a single gene in a cultivar to achieve durable resistance whereas if hypothesis 1 is correct then *Lr22a* must be used in combination with other *R* genes to breed for durable resistance.

Also, the cloned gene can be used to identify its homoeologs and orthologues in wheat and other cereals to identify the *R* gene specificity-determining residues. Identification of specificity-determining residues would facilitate precision engineering solutions whereby a non-functional allele in a susceptible cultivar is changed into a functional allele. For example, *Lr22a* allele mining in 25 hexaploid wheat genotypes without *Lr22a* resistance revealed two amino acids in the N terminus that were unique to Lr22a. Moreover, allele mining in 45 *Ae. tauschii* accessions (as *Ae. tauschii* is a donor of *Lr22a*), revealed 3 interesting *Lr22a* alleles

(Ackermann, 2018). One of the *Ae. tauschii* accession, AE#45, had an allele (allele 9) whose protein sequence differed from *Lr22a* by only 1 amino acid and showed complete resistance against a mixture of 16 leaf rust isolates. Another two *Ae. tauschii* accessions, AE#33 and AE#15, had the same *Lr22a* allele (allele 5) and the protein sequence differed from *Lr22a* and allele 5 by 22 and 21 amino acids, respectively, with the majority of the changes present in the LRR domain. Allele 5 interestingly produced a different resistance phenotype, partial in one accession (AE#33) and complete in other *Ae. tauschii* accession (AE#15), which might indicate genotype-by-genotype interaction. However, it is difficult to conclude whether the resistance observed in these *Ae. tauschii* accessions is *Lr22a* mediated or by some other R gene in the background. However, positive associations between disease resistance and polymorphic residues in NLRs can be tested by synthesizing gene variants aimed at rendering an *R* gene non-functional, or a non-functional allele functional. The effect of the mutation on function can be examined through stable transgenics or by virus-induced gene silencing (VIGS). In VIGS, viruses trigger host defense machinery related to post-transcriptional gene silencing, where double stranded RNA is converted to short interfering RNAs. Gene of interest in introduced into the virus and the recombinant virus triggers the host-defense response. Both the virus genome and the endogenous mRNAs homologous to the inserted target sequence become the targets for degradation (Ma, Yan, Huang, Chen, & Zhao, 2012). Silencing which is initiated by VIGS, spreads systematically along with the siRNA. This method can be used to knock out any target gene if a suitable vector is present for the plant species under inverstigation (Ma et al., 2012).

To conclude, understanding of molecular mechanisms will promote designing of new strategies for generating the next-generation crops resistant to multiple pathogens, for example, to engineer a *Lr* gene pyramid or generate ‘multi-lines’ that would provide more durable resistance by delaying the emergence of resistance-breaking pathogen strains (Brunner et al., 2012; Dangl et

al., 2013) and to overcome environmental stresses to reduce the dependence on pesticides and other chemicals.

References

Ackermann, P. M. (2018). Genomic and biochemical analysis of the *Lr22a* leaf rust locus in wheat. Master's thesis. *University of Zurich*.

Aguileta, G., Refregier, G., Yockteng, R., Fournier, E., & Giraud, T. (2009). Rapidly evolving genes in pathogens: Methods for detecting positive selection and examples among fungi, bacteria, viruses and protists Discussion. *Infect Genet Evol, 9*(4), 656-670. doi:10.1016/j.meegid.2009.03.010

Akhunov, E. D., Akhunova, A. R., Anderson, O. D., Anderson, J. A., Blake, N., Clegg, M. T., . . . Dvorak, J. (2010). Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genom, 11*(702). doi:70210.1186/1471-2164-11-702

Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet, 12*(5), 363-376. doi:10.1038/nrg2958

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Arora, S., Singh, N., Kaur, S., Bains, N. S., Uauy, C., Poland, J., & Chhuneja, P. (2017). Genome-Wide Association Study of Grain Architecture in Wild Wheat *Aegilops tauschii*. *Front Plant Sci, 8*(886). doi:10.3389/fpls.2017.00886

Arora, S., Steuernagel, B., Long, Y., Matny, O., Johnson, R., Enk, J., . . . Wulff, B. B. H. (2018). Resistance gene discovery and cloning by sequence capture and association genetics. *bioRxiv*. doi:10.1101/248146

Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., . . . Distelfeld, A. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science, 357*(6346), 93-97. doi:10.1126/science.aan0032

Axtell, M. J., Chisholm, S. T., Dahlbeck, D., & Staskawicz, B. J. (2003). Genetic and molecular evidence that the *Pseudomonas syringae* type III effector protein AvrRpt2 is a cysteine protease. *Mol Microbiol, 49*(6), 1537-1546. doi:10.1046/j.1365-2958.2003.03666

Ay, F., & Noble, W. S. (2015). Analysis methods for studying the 3D architecture of the genome. *Genome Biol, 16*(183). doi:10.1186/s13059-015-0745-7

Baggs, E., Dagdas, G., & Krasileva, K. V. (2017). NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. *Curr Opin Plant Biol, 38*, 59-67. doi:10.1016/j.pbi.2017.04.012

Beddow, J. M., Pardey, P. G., Chai, Y., Hurley, T. M., Kriticos, D. J., Braun, H. J., . . . Yonow, T. (2015). Research investment implications of shifts in the global geography of wheat stripe rust. *Nat Plants, 1*(10). doi:1513210.1038/Nplants.2015.132

Belkhadir, Y., Nimchuk, Z., Hubert, D. A., Mackey, D., & Dangl, J. L. (2004). Arabidopsis RIN4 negatively regulates disease resistance mediated by RPS2 and RPM1 downstream or independent of the NDR1 signal modulator and is not required for the virulence functions of bacterial type III effectors AvrRpt2 or AvrRpm1. *Plant Cell, 16*(10), 2822-2835. doi:10.1105/tpc.104.024117

Bennetzen, J. L. (2007). Patterns in grass genome evolution. *Curr Opin Plant Biol, 10*(2), 176-181. doi:10.1016/j.pbi.2007.01.010

Berkman, P. J., Skarshewski, A., Lorenc, M. T., Lai, K. T., Duran, C., Ling, E. Y. S., . . . Edwards, D. (2011). Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol J, 9*(7), 768-775. doi:10.1111/j.1467-7652.2010.00587

Berkman, P. J., Skarshewski, A., Manoli, S., Lorenc, M. T., Stiller, J., Smits, L., . . . Edwards, D. (2012). Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor Appl*

Genet, 124(3), 423-432. doi:10.1007/s00122-011-1717-2

Bieri, S., Mauch, S., Shen, Q. H., Peart, J., Devoto, A., Casais, C., . . . Schulze-Lefert, P. (2004). RAR1 positively controls steady state levels of barley MLA resistance proteins and enables sufficient MLA6 accumulation for effective resistance. *Plant Cell, 16*(12), 3480-3495. doi:10.1105/tpc.104.026682

Bisgrove, S. R., Simonich, M. T., Smith, N. M., Sattler, A., & Innes, R. W. (1994) . A disease resistance gene in *Arabidopsis* with specificity for two different pathogen avirulence genes. *Plant Cell, 6*(7), 927-933. doi:10.1105/tpc.6.7.927

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bolton, M. D., Kolmer, J. A., & Garvin, D. F. (2008). Wheat leaf rust caused by *Puccinia triticina*. *Mol Plant Pathol, 9*(5), 563-575. doi:10.1111/J.1364-3703.2008.00487

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L. A., D'Amore, R., Allen, A. M., . . . Hall, N. (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature, 491*(7426), 705-710. doi:10.1038/nature11650

Brunner, S., Stirnweis, D., Diaz Quijano, C., Buesing, G., Herren, G., Parlange, F., . . . Keller, B. (2012). Transgenic *Pm3* multilines of wheat show increased powdery mildew resistance in the field. *Plant Biotechnol J, 10*(4), 398-409. doi:10.1111/j.1467-7652.2011.00670

Buchmann, J. P., Matsumoto, T., Stein, N., Keller, B., & Wicker, T. (2012). Inter-species sequence comparison of *Brachypodium* reveals how transposon activity corrodes genome colinearity. *Plant J, 71*(4), 550-563. doi:10.1111/j.1365-313X.2012.05007.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol, 31*, 1119-1125. doi:10.1038/nbt.2727

Buschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., . . . Schulze-Lefert, P. (1997). The barley *Mlo* gene: a novel control element of plant pathogen resistance. *Cell, 88*(5), 695-705.

Cai, X., & Xu, S. S. (2007). Meiosis-driven genome variation in plants. *Curr Genomics, 8*(3), 151-161.

Cavanagh, C. R., Chao, S., Wang, S., Huang, B. E., Stephen, S., Kiani, S., . . . Akhunov, E. (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc Natl Acad Sci U S A, 110*(20), 8057-8062. doi:10.1073/pnas.1217133110

Cesari, S., Thilliez, G., Ribot, C., Chalvon, V., Michel, C., Jauneau, A., . . . Kroj, T. (2013). The rice resistance protein pair *RGA4/RGA5* recognizes the *Magnaporthe oryzae* effectors AVR-Pia and AVR1-CO39 by direct binding. *Plant Cell, 25*(4), 1463-1481. doi:10.1105/tpc.112.107201

Chakraborty, S., & Newton, A. C. (2011). Climate change, plant diseases and food security: an overview. *Plant Pathol, 60*(1), 2-14. doi:10.1111/j.1365-3059.2010.02411

Chantret, N., Salse, J., Sabot, F., Rahman, S., Bellec, A., Laubin, B., . . . Chalhoub, B. (2005). Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *Plant Cell, 17*(4), 1033-1045. doi:10.1105/tpc.104.029181

Chapman, J. A., Ho, I., Sunkara, S., Luo, S., Schroth, G. P., & Rokhsar, D. S. (2011). Meraculous: de novo genome assembly with short paired-end reads. *PLoS One, 6*(8), e23501. doi:10.1371/journal.pone.0023501

Chapman, J. A., Mascher, M., Buluc, A., Barry, K., Georganas, E., Session, A., . . . Rokhsar, D. S. (2015). A whole-genome shotgun approach for assembling and anchoring the

hexaploid bread wheat genome. *Genome Biol, 16*(26). doi:10.1186/s13059-015-0582-8

Chen, X. M. (2005). Epidemiology and control of stripe rust [*Puccinia striiformis* f. sp *tritici*] on wheat. *Can J Plant Pathol, 27*(3), 314-337. doi:10.1146/annurev.phyto.38.1.491

Chia, J. M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., . . . Ware, D. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet, 44*(7), 803-807. doi:10.1038/ng.2313

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., . . . Feuillet, C. (2014). Structural and functional partitioning of bread wheat chromosome 3B. *Science, 345*(6194), 1249721. doi:10.1126/science.1249721

Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakottil, G., Wright, J., . . . Clark, M. D. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res, 27*(5), 885-896. doi:10.1101/gr.217117.116

Cloutier, S., McCallum, B. D., Loutre, C., Banks, T. W., Wicker, T., Feuillet, C., . . . Jordan, M. C. (2007). Leaf rust resistance gene *Lr1*, isolated from bread wheat (*Triticum aestivum* L.) is a member of the large psr567 gene family. *Plant Mol Biol, 65*(1-2), 93-106. doi:10.1007/s11103-007-9201-8

Cook, D. E., Mesarich, C. H., & Thomma, B. P. H. J. (2015). Understanding Plant Immunity as a Surveillance System to Detect Invasion. *Annu Rev Phytopathol, 53*, 541-563. doi:10.1146/annurev-phyto-080614-120114

Dangl, J. L., & Jones, J. D. G. (2001). Plant pathogens and integrated defence responses to infection. *Nature, 411*(6839), 826-833. Doi:10.1038/35081161

Dangl, J. L., Horvath, D. M., & Staskawicz, B. J. (2013). Pivoting the plant immune system from dissection to deployment. *Science, 341*(6147), 746-751. doi:10.1126/science.1236011

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., & Hahn, M. W. (2014). Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol, 10*(12), e1003998. doi:10.1371/journal.pcbi.1003998

Dodds, P. N., & Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat Rev Genet, 11*(8), 539-548. doi:10.1038/nrg2812

Dodds, P. N., Lawrence, G. J., & Ellis, J. G. (2001). Six amino acid changes confined to the leucine-rich repeat beta-strand/beta-turn motif determine the difference between the P and P2 rust resistance specificities in flax. *Plant Cell, 13*(1), 163-178. doi: 10.1105/tpc.13.1.163

Dolezel, J., Vrana, J., Safar, J., Bartos, J., Kubalakova, M., & Simkova, H. (2012). Chromosomes in the flow to simplify genome analysis. *Funct Integr Genom, 12*(3), 397-416. doi:10.1007/s10142-012-0293-0

Dreisigacker, S., Kishii, M., Lage, J., & Warburton, M. (2008). Use of synthetic hexaploid wheat to increase diversity for CIMMYT bread wheat improvement. *Aust J Agricul Res, 59*(5), 413-420. doi:10.1071/Ar07225

Dvorak, J., Deal, K. R., Luo, M. C., You, F. M., von Borstel, K., & Dehghani, H. (2012). The origin of spelt and free-threshing hexaploid wheat. *J Hered, 103*(3), 426-441. doi:10.1093/jhered/esr152

Dyck P. L., Kerber E. R. (1970). Inheritance in hexaploid wheat of adult-plant leaf rust resistance derived from *Aegilops squarrosa*. *Can J Genet Cytol, 12*(1), 175-180. doi:10.1139/g70-025

Ellis, J. G., Dodds, P. N., & Lawrence, G. J. (2007). Flax rust resistance gene specificity is based on direct resistance-avirulence protein interactions. *Annu Rev Phytopathol, 45*, 289-306. doi:10.1146/annurev.phyto.45.062806.094331

Ellis, J. G., Lagudah, E. S., Spielmeyer, W., & Dodds, P. N. (2014). The past, present and future of breeding rust resistant wheat. *Front Plant Sci, 5*:641. doi: 10.3389/fpls.2014.00641

Endo, T. R., & Gill, B. S. (1996). The deletion stocks of common wheat. *J Hered, 87*(4), 295-307. doi:10.1093/oxfordjournals.jhered.a023003

FAO (2017). The Food and Agriculture Organization (FAO).http://www.fao.org/faostat/en/.

FAO (2011). The state of the world's land and water resources for food and agriculture (SOLAW) - Managing systems at risk, (Food and Agriculture Organization of the United Nations, Rome and Earthscan, London).

Feldman, M., & Levy, A. A. (2005). Allopolyploidy-A shaping force in the evolution of wheat genomes. *Cytogenet Genome Res, 109*(1-3), 250-258. doi:10.1159/000082407

Feuillet, C., Travella, S., Stein, N., Albar, L., Nublat, A., & Keller, B. (2003). Map-based isolation of the leaf rust disease resistance gene *Lr10* from the hexaploid wheat (*Triticum aestivum* L.) genome. *Proc Natl Acad Sci U S A, 100*(25), 15253-15258. doi:10.1073/pnas.2435133100

Figueroa, M., Hammond-Kosack, K. E., & Solomon, P. S. (2017). A review of wheat diseases-a field perspective. *Mol Plant Pathol*. doi:10.1111/mpp.12618

Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L., & Gurr, S. J. (2012). Emerging fungal threats to animal, plant and ecosystem health. *Nature, 484*(7393), 186-194. doi:10.1038/nature10947

Fishman-Lobell Jacqueline, R. N., Haber J. E. (1992). Two alternative pathways of double-strand break repair that are kinetically separable and independently modulated. *Mol Cell Biol*, *12*(3) 1292-1303.

Fu, D., Uauy, C., Distelfeld, A., Blechl, A., Epstein, L., Chen, X., . . . Dubcovsky, J. (2009). A kinase-START gene confers temperature-dependent resistance to wheat stripe rust. *Science, 323*(5919), 1357-1360. doi:10.1126/science.1166289

Fukuoka, S., & Okuno, K. (2001). QTL analysis and mapping of pi21, a recessive gene for field resistance to rice blast in Japanese upland rice. *Theor Appl Genet, 103*(2-3), 185-190.

Fukuoka, S., Saka, N., Koga, H., Ono, K., Shimizu, T., Ebana, K., . . . Yano, M. (2009). Loss of function of a proline-containing protein confers durable disease resistance in rice. *Science, 325*(5943), 998-1001. doi:10.1126/science.1175550

Gao, Z., Chung, E. H., Eitas, T. K., & Dangl, J. L. (2011). Plant intracellular innate immune receptor Resistance to *Pseudomonas syringae* pv. *maculicola 1* (RPM1) is activated at, and functions on, the plasma membrane. *Proc Natl Acad Sci U S A, 108*(18), 7619-7624. doi:10.1073/pnas.1104410108

Gardiner, L. J., Bansept-Basler, P., Olohan, L., Joynson, R., Brenchley, R., Hall, N., . . . Hall, A. (2016). Mapping-by-sequencing in complex polyploid genomes using genic sequence capture: a case study to map yellow rust resistance in hexaploid wheat. *Plant J, 87*(4), 403-419. doi:10.1111/tpj.13204

Giorgi, D., Farina, A., Grosso, V., Gennaro, A., Ceoloni, C., & Lucretti, S. (2013). FISHIS: fluorescence in situ hybridization in suspension and chromosome flow sorting made easy. *PLoS One, 8*(2), e57994. doi:10.1371/journal.pone.0057994

Gottlieb, A., Muller, H. G., Massa, A. N., Wanjugi, H., Deal, K. R., You, F. M., . . . Dvorak, J. (2013). Insular organization of gene space in grass genomes. *PLoS One, 8*(1), e54101. doi:10.1371/journal.pone.0054101

Grant, M. R., Godiard, L., Straube, E., Ashfield, T., Lewald, J., Sattler, A., . . . Dangl, J. L. (1995). Structure of the *Arabidopsis* RPM1 gene enabling dual specificity disease resistance. *Science, 269*(5225), 843-846.

Greenwood, T. A., Rana, B. K., & Schork, N. J. (2004). Human haplotype block sizes are negatively correlated with recombination rates. *Genome Res, 14*(7), 1358-1361. doi:10.1101/gr.1540404

Gremme, G., Brendel, V., Sparks, M. E., & Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. *47*(15), 965-978. doi: 10.1016/j.infsof.2005.09.005

Helft, L., Reddy, V., Chen, X., Koller, T., Federici, L., Fernandez-Recio, J., . . . Bent, A. (2011). LRR conservation mapping to predict functional sites within protein leucine-rich repeat domains. *PLoS One, 6*(7), e21614. doi:10.1371/journal.pone.0021614

Hiebert, C. W., Thomas, J. B., Somers, D. J., McCallum, B. D., & Fox, S. L. (2007). Microsatellite mapping of adult-plant leaf rust resistance gene *Lr22a* in wheat. *Theor Appl Genet, 115*(6), 877-884. doi:10.1007/s00122-007-0604-3

Hirsch, C. N., Hirsch, C. D., Brohammer, A. B., Bowman, M. J., Soifer, I., Barad, O., . . . Mikel, M. A. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *Plant Cell, 28*(11), 2700-2714. doi:10.1105/tpc.16.00353

Hu, K., Cao, J., Zhang, J., Xia, F., Ke, Y., Zhang, H., . . . Wang, S. (2017). Improvement of multiple agronomic traits by a disease resistance gene via cell wall reinforcement. *Nat Plants, 3*, 17009. doi:10.1038/nplants.2017.9

Huang, L., Brooks, S. A., Li, W., Fellers, J. P., Trick, H. N., & Gill, B. S. (2003). Map-based cloning of leaf rust resistance gene *Lr21* from the large and polyploid genome of bread wheat. *Genetics, 164*(2), 655-664.

Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., & Gornicki, P. (2002). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc Natl Acad Sci U S A, 99*(12), 8133-8138. doi:10.1073/pnas.072223799

Huerta-Espino, J., Singh, R. P., German, S., McCallum, B. D., Park, R. F., Chen, W. Q., . . . Goyeau, H. (2011). Global status of wheat leaf rust caused by *Puccinia triticina*. *Euphytica, 179*(1), 143-160. doi:10.1007/s10681-011-0361

Hurni, S., Scheuermann, D., Krattinger, S. G., Kessel, B., Wicker, T., Herren, G., . . . Keller, B. (2015). The maize disease resistance gene *Htn1* against northern corn leaf blight encodes a wall-associated receptor-like kinase. *Proc Natl Acad Sci U S A, 112*(28), 8780-8785. doi:10.1073/pnas.1502522112

International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature, 463*(7282), 763-768. doi:10.1038/nature08747

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature, 436*(7052), 793-800. doi:10.1038/nature03895

International Wheat Genome Sequencing Consortium (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science, 345*(6194), 1251788. doi:10.1126/science.1251788

International Wheat Genome Sequencing Consortium (2018). Shifting the limits in wheat research and breeding through a fully annotated and anchored reference genome sequence *Science, under review*.

Isidore, E., Scherrer, B., Chalhoub, B., Feuillet, C., & Keller, B. (2005). Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been

maintained in species of three different ploidy levels. *Genome Res, 15*(4), 526-536. doi:10.1101/gr.3131005

Jacob, F., Vernaldi, S., & Maekawa, T. (2013). Evolution and Conservation of Plant NLR Functions. *Front Immunol, 4*, 297. doi:10.3389/fimmu.2013.00297

Jarvis, D. E., Ho, Y. S., Lightfoot, D. J., Schmockel, S. M., Li, B., Borm, T. J., . . . Tester, M. (2017). The genome of *Chenopodium quinoa*. *Nature, 542*(7641), 307-312. doi:10.1038/nature21370

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., . . . Wang, J. (2013). *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature, 496*(7443), 91-95. doi:10.1038/nature12028

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., . . . Ware, D. (2017). Improved maize reference genome with single-molecule technologies. *Nature, 546*(7659), 524-527. doi:10.1038/nature22971

Jin, Y. (2011). Role of Berberis spp. as alternate hosts in generating new races of *Puccinia graminis* and *P. striiformis*. *Euphytica, 179*(1), 105-108. doi:10.1007/s10681-010-0328-3

Johnson, R. (1984). A critical analysis of durable resistance. *Ann. Rev. Phytopathol., 22*, 309-330.

Johnson, R. (1988). Durable resistance to yellow (stripe) rust in wheat and its implications in plant breeding. in Breeding Strategies for Resistance to the Rusts of Wheat, eds N.W. Simmonds and S.Rajaram (Mexico:CIMMYT).

Jones, J. D. G., & Dangl, J. L. (2006). The plant immune system. *Nature, 444*(7117), 323-329. doi:10.1038/nature05286

Jordan, K. W., Wang, S., Lun, Y., Gardiner, L. J., MacLachlan, R., Hucl, P., . . . Akhunov, E. (2015). A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol, 16*, 48. doi:10.1186/s13059-015-0606-4

Jupe, F., Witek, K., Verweij, W., Sliwka, J., Pritchard, L., Etherington, G. J., . . . Jones, J. D. (2013). Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB- LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant J, 76*(3), 530-544. doi:10.1111/tpj.12307

Kanzaki, H., Yoshida, K., Saitoh, H., Fujisaki, K., Hirabuchi, A., Alaux, L., . . . Terauchi, R. (2012). Arms race co-evolution of *Magnaporthe oryzae* AVR-Pik and rice Pik genes driven by their physical interactions. *Plant Journal, 72*(6), 894-907. doi:10.1111/j.1365-313X.2012.05110

Kasmia, F. E., & Nishimurab, M. T. (2016). Structural insights into plant NLR immune receptor function. *Proc Natl Acad Sci U S A., 113*(45), 12619–12621.

Keller, B., Feuillet, C., & Yahiaoui, N. (2005). Map-based isolation of disease resistance genes from bread wheat: cloning in a supersize genome. *Genet Res, 85*(2), 93-100. doi:10.1017/S0016672305007391

Kim, M. G., Geng, X. Q., Lee, S. Y., & Mackey, D. (2009). The *Pseudomonas syringae* type III effector AvrRpm1 induces significant defenses by activating the *Arabidopsis* nucleotide-binding leucine-rich repeat protein RPS2. *Plant Journal, 57*(4), 645-653. doi:10.1111/j.1365-313X.2008.03716

Kolmer, J. (2013). Leaf Rust of Wheat: Pathogen Biology, Variation and Host Resistance. *Forests, 4*(1), 70-84. doi:10.3390/f4010070

Kolmer, J. A. (1997). Virulence in *Puccinia recondita* f. sp. *tritici* isolates from Canada to genes for adult-plant resistance to wheat leaf rust. *Plant Dis, 81*, 267-271.

Kolmer, J. A., Jin, Y., & Long, D. L. (2007). Wheat leaf and stem rust in the United States. *Aust J Agricul Res, 58*(6), 631-638. doi:10.1071/Ar07057

Krattinger, S. G., & Keller, B. (2016). Molecular genetics and evolution of disease resistance in cereals. *New Phytol, 212*(2), 320-332. doi:10.1111/nph.14097

Krattinger, S. G., Lagudah, E. S., Spielmeyer, W., Singh, R. P., Huerta-Espino, J., McFadden, H., . . . Keller, B. (2009). A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science, 323*(5919), 1360-1363. doi:10.1126/science.1166453

Krattinger, S., Wicker, T., & Keller, B. (2007). Map-Based Cloning of Genes in Triticeae (Wheat and Barley). in *Genetics and Genomics of the Triticeae, 7*, 337-357. doi:10.1007/978-0-387-77489-3_12

Kubalakova, M., Valarik, M., Barto, J., Vrana, J., Cihalikova, J., Molnar-Lang, M., & Dolezel, J. (2003). Analysis and sorting of rye (*Secale cereale* L.) chromosomes using flow cytometry. *Genome, 46*(5), 893-905. doi:10.1139/g03-054

Kubalakova, M., Vrana, J., Cihalikova, J., Simkova, H., & Dolezel, J. (2002). Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor Appl Genet, 104*(8), 1362-1372. doi:10.1007/s00122-002-0888-2

Leitch, A. R., & Leitch, I. J. (2008). Genomic plasticity and the diversity of polyploid plants. *Science, 320*(5875), 481-483. doi:10.1126/science.1153585

Leonard, K. J., & Szabo, L. J. (2005). Stem rust of small grains and grasses caused by *Puccinia graminis*. *Mol Plant Pathol, 6*(2), 99-111. doi:10.1111/J.1364.3703.2004.00273

Lewis, C. M., Persoons, A., Bebber, D. P., Kigathi, R. N., Maintz, J., Findlay, K., . . . Saunders, D. G. O. (2018). Potential for re-emergence of wheat stem rust in the United Kingdom. *Comm Biol, 1*(13). doi:10.1038/s42003-018-0013

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science, 326*(5950), 289-293. doi:10.1126/science.1181369

Ling, H. Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., . . . Liang, C. (2018). Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature, 557*(7705), 424-428. doi:10.1038/s41586-018-0108-0

Ling, H. Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., . . . Wang, J. (2013). Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature, 496*(7443), 87-90. doi:10.1038/nature11997

Liu, M., Stiller, J., Holusova, K., Vrana, J., Liu, D., Dolezel, J., & Liu, C. (2016). Chromosome-specific sequencing reveals an extensive dispensable genome component in wheat. *Sci Rep, 6*, 36398. doi:10.1038/srep36398

Liu, Q., Liu, H., Gong, Y., Tao, Y., Jiang, L., Zuo, W., . . . Xu, M. (2017). An atypical thioredoxin imparts early resistance to sugarcane mosaic virus in maize. *Mol Plant, 10*(3), 483-497. doi:10.1016/j.molp.2017.02.002

Loutre, C., Wicker, T., Travella, S., Galli, P., Scofield, S., Fahima, T., . . . Keller, B. (2009). Two different CC-NBS-LRR genes are required for *Lr10*-mediated leaf rust resistance in tetraploid and hexaploid wheat. *Plant J, 60*(6), 1043-1054. doi:10.1111/j.1365-313X.2009.04024

Luo, M. C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., . . . Dvorak, J. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature, 551*(7681), 498-502. doi:10.1038/nature24486

Luo, M. C., Gu, Y. Q., You, F. M., Deal, K. R., Ma, Y., Hu, Y., . . . Dvorak, J. (2013). A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *Proc Natl Acad Sci U S A, 110*(19), 7940-7945. doi:10.1073/pnas.1219082110

Ma, J., & Bennetzen, J. L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A, 101*(34), 12404-12410. doi:10.1073/pnas.0403715101

Ma, M., Yan, Y., Huang, L., Chen, M. S., & Zhao, H. X. (2012). Virus-induced gene-silencing in wheat spikes and grains and its application in functional analysis of HMW-GS-encoding genes. *BMC Plant Biol, 12*. doi:14110.1186/1471-2229-12-141

Mackey, D., Holt, B. F., Wiig, A., & Dangl, J. L. (2002). RIN4 interacts with *Pseudomonas syringae* type III effector molecules and is required for RPM1-mediated resistance in *Arabidopsis*. *Cell, 108*(6), 743-754. doi:10.1016/S0092-8674(02)00661

Mago, R., Tabe, L., Vautrin, S., Simkova, H., Kubalakova, M., Upadhyaya, N., . . . Spielmeyer, W. (2014). Major haplotype divergence including multiple germin-like protein genes, at the wheat Sr2 adult plant stem rust resistance locus. *BMC Plant Biol, 14*, 379. doi:10.1186/s12870-014-0379

Mago, R., Zhang, P., Vautrin, S., Simkova, H., Bansal, U., Luo, M. C., . . . Dodds, P. N. (2015). The wheat *Sr50* gene reveals rich diversity at a cereal disease resistance locus. *Nat Plants, 1*(12). doi:1518610.1038/Nplants.2015.186

Maqbool, A., Saitoh, H., Franceschetti, M., Stevenson, C. E., Uemura, A., Kanzaki, H., . . . Banfield, M. J. (2015). Structural basis of pathogen recognition by an integrated HMA domain in a plant NLR immune receptor. *Elife, 4*. doi:10.7554/eLife.08709

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., . . . Stein, N. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature, 544*(7651), 426-433. doi:10.1038/nature22043

McCallum, B. D., Hiebert, C. W., Cloutier, S., Bakkeren, G., Rosa, S. B., Humphreys, D. G., . . . Wang, X. B. (2016). A review of wheat leaf rust research and the development of resistant cultivars in Canada. *Can J Plant Pathol, 38*(1), 1-18. doi:10.1080/07060661.2016.1145598

McCallum, B. D., Seto-Goh, P., & Xue, A. (2013). Physiologic specialization ofPuccinia triticina, the causal agent of wheat leaf rust, in Canada in 2009. *Can J Plant Pathol, 35*(3), 338-345. doi:10.1080/07060661.2013.810669

Mcfadden, E. S., & Sears, E. R. (1944). The Artificial Synthesis of Triticum-Spelta. *Rec Genet Soc Am, 13*, 26-27.

McIntosh, R., Wellings, C., & Park, R. (1995). Wheat Rusts:An Atlas of Resistance Genes. Melbourne:CSIRO Publishing.

Mendgen, K., & Hahn, M. (2002). Plant infection and the establishment of fungal biotrophy. *Trends Plant Sci, 7*(8), 352-356. doi:10.1016/S1360-1385(02)02297-5

Middleton, C. P., Stein, N., Keller, B., Kilian, B., & Wicker, T. (2013). Comparative analysis of genome composition in Triticeae reveals strong variation in transposable element dynamics and nucleotide diversity. *Plant J, 73*(2), 347-356. doi:10.1111/tpj.12048

Moll, K. M., Zhou, P., Ramaraj, T., Fajardo, D., Devitt, N. P., Sadowsky, M. J., . . . Mudge, J. (2017). Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula. *BMC Genom, 18*, 578. doi:10.1186/s12864-017-3971-4

Molnár-Láng M. , Ceoloni C., Doležel J. (2015). Alien introgression in wheat: Cytogenetics, molecular biology, and genomics. *Springer International Publishing*, 385 doi:10.1007/978-3-319-23494-6

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K., . . . Edwards, D. (2017). The pangenome of hexaploid bread wheat. *Plant J, 90*(5), 1007-1013. doi:10.1111/tpj.13515

Moore, J. W., Herrera-Foessel, S., Lan, C. X., Schnippenkoetter, W., Ayliffe, M., Huerta-Espino, J., . . . Lagudah, E. (2015). A recently evolved hexose transporter variant

confers resistance to multiple pathogens in wheat. *Nat Genet, 47*(12), 1494-1498. doi:10.1038/ng.3439

Moullet Odile , S. A. (2014). Maintaining the efficiency of MAS method in cereals while reducing the costs. *J. Plant Breed. Genet., 2*(2), 97–100.

Munoz-Amatriain, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., . . . Stein, N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol, 14*(6), R58. doi:10.1186/gb-2013-14-6-r58

Nimchuk, Z., Eulgem, T., Holt, B. E., & Dangl, J. L. (2003). Recognition and response in the plant immune system. *Annu Rev Genet, 37*, 579-609. doi:10.1146/annurev.genet.37.110801.142628

Oerke, E. C. (2006). Crop losses to pests. *J Agricult Sci, 144*, 31-43. doi:10.1017/S0021859605005708

Pardey, P. G., Beddow, J. M., Kriticos, D. J., Hurley, T. M., Park, R. F., Duveiller, E., . . . Hodson, D. (2013). Agriculture. Right-sizing stem-rust research. *Science, 340*(6129), 147-148. doi:10.1126/science.122970

Parniske, M., HammondKosack, K. E., Golstein, C., Thomas, C. M., Jones, D. A., Harrison, K., . . . Jones, J. D. G. (1997). Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell, 91*(6), 821-832. doi:10.1016/S0092-8674(00)80470-5

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., . . . Rokhsar, D. S. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature, 457*(7229), 551-556. doi:10.1038/nature07723

Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., . . . Feuillet, C. (2008). A physical map of the 1-gigabase bread wheat chromosome 3B. *Science, 322*(5898), 101-104. doi:10.1126/science.1161847

Pearce, S., Zhu, J., Boldizsar, A., Vagujfalvi, A., Burke, A., Garland-Campbell, K., . . . Dubcovsky, J. (2013). Large deletions in the CBF gene cluster at the *Fr-B2* locus are associated with reduced frost tolerance in wheat. *Theor Appl Genet, 126*(11), 2683-2697. doi:10.1007/s00122-013-2165

Peng, J. H. H., Sun, D. F., & Nevo, E. (2011). Domestication evolution, genetics and genomics in wheat. *Mol Breeding, 28*(3), 281-301. doi:10.1007/s11032-011-9608-4

Periyannan S., Moore J., Ayliffe M., Bansal U., Wang X., Huang L ., . . . Lagudah E. (2013). The Gene *Sr33*, an Ortholog of Barley Mla Genes, Encodes Resistance to Wheat Stem Rust Race Ug99. *Science*. 6;341(6147):786-788.

Pfeiffer, P., Goedecke, W., & Obe, G. (2000). Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis, 15*(4), 289-302. doi:10.1093/mutage/15.4.289

Pretorius, Z. (2000). Detection of virulence to wheat stem rust resistance gene Sr31 in Puccinia graminis. f.sp. tritici in Uganda. *Plant Dis, 84*, 203. doi:10.1094/PDIS.2000.84.2.203B Pretorius, Z. A., Rijkenberg, F. H. J., & Wilcoxson, R. D. (1987). Characterization of adult-plant resistance to leaf rust of wheat conferred by the gene *Lr22a*. *Plant Dis, 71*, 542-545.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One, 5*(3), e9490. doi:10.1371/journal.pone.0009490

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., . . . Green, R. E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res, 26*(3), 342-350. doi:10.1101/gr.193474.115

Ramakrishna, W., Emberton, J., Ogden, M., SanMiguel, P., & Bennetzen, J. L. (2002). Structural analysis of the maize rp1 complex reveals numerous sites and unexpected mechanisms of local rearrangement. *Plant Cell, 14*(12), 3213-3223.

Ramirez-Gonzalez, R. H., Segovia, V., Bird, N., Fenwick, P., Holdgate, S., Berry, S., . . . Uauy, C. (2015). RNA-Seq bulked segregant analysis enables the identification of high-resolution genetic markers for breeding in hexaploid wheat. *Plant Biotechnol J, 13*(5), 613-624. doi:10.1111/pbi.12281

Rawat, N., Pumphrey, M. O., Liu, S., Zhang, X., Tiwari, V. K., Ando, K., . . . Gill, B. S. (2016). Wheat *Fhb1* encodes a chimeric lectin with agglutinin domains and a pore-forming toxin-like domain conferring resistance to Fusarium head blight. *Nat Genet, 48*(12), 1576-1580. doi:10.1038/ng.3706

Retief, J. D. (2000). Phylogenetic analysis using PHYLIP. *Methods Mol Biol, 132*, 243-258.

Robberecht C, V. T., Zamani Esteki M, Nowakowska BA, Vermeesch JR. (2013). Non-allelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res, 23*, 411-418. doi:10.1101/gr.145631.112

Roelfs, A. P. (1984). Race specificity and methods of study. In A. P. Roelfs & W. R. Bushnell (Eds.), *The Cereal Rusts Vol. I; Origins, specificity, structure, and physiology*. Orlando: Academic Press.

Roffler, S., & Wicker, T. (2015). Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons. *Mob DNA, 6*(8). doi:10.1186/s13100-015-0040

Roffler, S., Menardo, F., & Wicker, T. (2015). The making of a genomic parasite—the Mothra family sheds light on the evolution of Helitrons in plants. *Mob. DNA, 6*(23). doi:10.1186/s13100-015-0054-4

Saari E.E., & Prescott J.M. (1985). World distribution in relation to economic losses. In: Roelfs AP, Bushnell WR (eds) The cereal rusts, vol 2, diseases, distribution, epidemiology, and control. *Academic Press, Orlando, Florida*, 259-298.

Safar, J., Simkova, H., Kubalakova, M., Cihalikova, J., Suchankova, P., Bartos, J., & Dolezel, J. (2010). Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet Genome Res, 129*(1-3), 211-223. doi:10.1159/000313072

Saintenac, C., Lee, W. S., Cambon, F., Rudd, J. J., King, R. C., Marande, W., . . . Kanyuka, K. (2018). Wheat receptor-kinase-like protein *Stb6* controls gene-for-gene resistance to fungal pathogen Zymoseptoria tritici. *Nat Genet, 50*(3), 368-374. doi:10.1038/s41588-018-0051

Saintenac, C., Zhang, W. J., Salcedo, A., Rouse, M. N., Trick, H. N., Akhunov, E., & Dubcovsky, J. (2013). Identification of Wheat Gene Sr35 That Confers Resistance to Ug99 Stem Rust Race Group. *Science, 341*(6147), 783-786. doi:10.1126/science.1239022

Salamini, F., H. Özkan, A., Brandolini, R., Schäfer-Pregl, & W. Martin. (2002). Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet., 3*, 429-441.

Salomon, S., & Puchta, H. (1998). Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. *Embo Journal, 17*(20), 6086-6095. doi:10.1093/emboj/17.20.6086

Sanchez-Martin, J., Steuernagel, B., Ghosh, S., Herren, G., Hurni, S., Adamski, N., . . . Wulff, B. B. (2016). Rapid gene isolation in barley and wheat by mutant chromosome sequencing. *Genome Biol, 17*(1), 221. doi:10.1186/s13059-016-1082-1

Sandhu, D., Gao, H. Y., Cianzio, S., & Bhattacharyya, M. K. (2004). Deletion of a disease resistance nucleotide-binding-site leucine-rich-repeat-like sequence is associated with the loss of the Phytophthora resistance gene Rps4 in soybean. *Genetics, 168*(4), 2157-2167. doi:10.1534/genetics.104.032037

Sarris, P. F., Cevik, V., Dagdas, G., Jones, J. D., & Krasileva, K. V. (2016). Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biol, 14*, 8. doi:10.1186/s12915-016-0228-7

Saxena, R. K., Edwards, D., & Varshney, R. K. (2014). Structural variations in plant genomes. *Brief Funct Genomics, 13*(4), 296-307. doi:10.1093/bfgp/elu016

Schmidt, M. H. W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., . . . Usadel, B. (2017). De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *Plant Cell, 29*(10), 2336-2348. doi:10.1105/tpc.17.00521

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., . . . Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science, 326*(5956), 1112-1115. doi:10.1126/science.1178534

Schwessinger, B. (2017). Fundamental wheat stripe rust research in the 21(st) century. *New Phytol, 213*(4), 1625-1631. doi:10.1111/nph.14159

Sears, E. R., & Sears, L. M. S. (1978). The Telocentric Chromosomes of Common Wheat. . *Proceedings of the 5th International Wheat Genetics Symposium, New Delhi*, 389-407.

Shatalina, M., Wicker, T., Buchmann, J. P., Oberhaensli, S., Simkova, H., Dolezel, J., & Keller, B. (2013). Genotype-specific SNP map based on whole chromosome 3B sequence information from wheat cultivars Arina and Forno. *Plant Biotechnol J, 11*(1), 23-32. doi:10.1111/pbi.12003

Shevelev, I. V., & Hubscher, U. (2002). The 3 '-5 ' exonucleases. *Nat Rev Mol Cell Biol, 3*(5), 364-375. doi:10.1038/nrm804

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., . . . Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol, 7*(539). doi:10.1038/msb.2011.75

Šimková, H., Číhalíková, J., Vrána, J., Lysák, M. A., & Doležel, J. (2003). Preparation of HMW DNA from plant nuclei and chromosomes isolated from root tips. *Biologia Plantarum, 46*, 369-373.

Simkova, H., Svensson, J. T., Condamine, P., Hribova, E., Suchankova, P., Bhat, P. R., . . . Dolezel, J. (2008). Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genom, 9*, 294. doi:10.1186/1471-2164-9-294

Singh, R. P., Herrera-Foessel, S., Huerta-Espino, J., Singh, S., Bhavani, S., Lan, C. X., & Basnet, B. R. (2014). Progress towards genetics and breeding for minor genes based resistance to ug99 and other rusts in cimmyt high-yielding spring wheat. *J Integr Agricult, 13*(2), 255-261. doi:10.1016/S2095-3119(13)60649-8

Singh, R. P., Hodson, D. P., Huerta-Espino, J., Jin, Y., Bhavani, S., Njau, P., . . . Govindan, V. (2011). The emergence of ug99 races of the stem rust fungus is a threat to world wheat production. *Annu Rev Phytopathol, 49*, 465-481. doi:10.1146/annurev-phyto-072910-095423

Singh, R. P., Hodson, D. P., Jin, Y., Lagudah, E. S., Ayliffe, M. A., Bhavani, S., . . . Hovmoller, M. S. (2015). Emergence and spread of new races of wheat stem rust fungus: continued threat to food security and prospects of genetic control. *Phytopathology, 105*(7), 872-884. doi:10.1094/PHYTO-01-15-0030-FI

Singla, J., Luthi, L., Wicker, T., Bansal, U., Krattinger, S. G., & Keller, B. (2017). Characterization of *Lr75*: a partial, broad-spectrum leaf rust resistance gene in wheat. *Theor Appl Genet, 130*(1), 1-12. doi:10.1007/s00122-016-2784-1

Sonnhammer, E. L., & Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene, 167*(1-2), GC1-10.

Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M., & Bergelson, J. (1999). Dynamics of disease resistance polymorphism at the Rpm1 locus of *Arabidopsis*. *Nature, 400*(6745), 667-671.

Stein, N., Feuillet, C., Wicker, T., Schlagenhauf, E., & Keller, B. (2000). Subgenome chromosome walking in wheat: a 450-kb physical contig in *Triticum monococcum* L. spans the *Lr10* resistance locus in hexaploid wheat (*Triticum aestivum* L.). *Proc Natl Acad Sci U S A, 97*(24), 13436-13441. doi:10.1073/pnas.230361597

Stein, N., Herren, G., & Keller, B. (2001). A new DNA extraction method for high-throughput marker analysis in a large-genome species such as *Triticum aestivum*. *Plant Breed, 120*, 354-356.

Steuernagel, B., Periyannan, S. K., Hernandez-Pinzon, I., Witek, K., Rouse, M. N., Yu, G., . . . Wulff, B. B. (2016). Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nat Biotechnol, 34*(6), 652-655. doi:10.1038/nbt.3543

Stirnweis, D., Milani, S. D., Brunner, S., Herren, G., Buchmann, G., Peditto, D., . . . Keller, B. (2014). Suppression among alleles encoding nucleotide-binding-leucine-rich repeat resistance proteins interferes with resistance in F1 hybrid and allele-pyramided wheat plants. *Plant J, 79*(6), 893-903. doi:10.1111/tpj.12592

Storici, F., Snipe, J. R., Chan, G. K., Gordenin, D. A., & Resnick, M. A. (2006). Conservative repair of a chromosomal double-strand break by single-strand DNA through two steps of annealing. *Mol Cell Biol, 26*(20), 7645-7657. doi:10.1128/Mcb.00672-06

Sudupak, M. A., Bennetzen, J. L., & Hulbert, S. H. (1993). Unequal exchange and meiotic instability of disease-resistance genes in the rp1 region of maize. *Genetics, 133*(1), 119-125.

Tanksley, S. D., & McCouch, S. R. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. *Science, 277*(5329), 1063-1066.

The 3,000 rice genomes project (2014). The 3,000 rice genomes project. *Gigascience, 3*, 7. doi:10.1186/2047-217X-3-7

Thind, A. K., Wicker, T., Simkova, H., Fossati, D., Moullet, O., Brabant, C., . . . Krattinger, S. G. (2017). Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. *Nat Biotechnol, 35*(8), 793-796. doi:10.1038/nbt.3877

Thomma, B. P. H. J., Nurnberger, T., & Joosten, M. H. A. J. (2011). Of PAMPs and Effectors: The blurred PTI-ETI dichotomy. *Plant Cell, 23*(1), 4-15. doi:10.1105/tpc.110.082602

Tilman, D., Balzer, C., Hill, J., & Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences of the United States of America, 108*(50), 20260-20264. doi:10.1073/pnas.1116437108

Trick, M., Adamski, N. M., Mugford, S. G., Jiang, C. C., Febrer, M., & Uauy, C. (2012). Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. *Bmc Plant Biol, 12*. doi:Artn 1410.1186/1471-2229-12-14

Uauy, C. (2017). Wheat genomics comes of age. *Curr Opin Plant Biol, 36*, 142-148. doi:10.1016/j.pbi.2017.01.007

Uauy, C., Brevis, J. C., Chen, X. M., Khan, I., Jackson, L., Chicaiza, O., . . . Dubcovsky, J. (2005). High-temperature adult-plant (HTAP) stripe rust resistance gene *Yr36* from *Triticum turgidum* ssp d*icoccoides* is closely linked to the grain protein content locus Gpc-B1. *Theor Appl Genet, 112*(1), 97-105. doi:10.1007/s00122-005-0109

Vale, F., Parlevliet, J., & Zambolim, L. (2001). Concepts in plant disease resistance. *Fitopatol. Bras., 26*, 577-589.

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., . . . Lander, E. S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*(39). doi:10.3791/1869

van der Biezen, E. A., & Jones, J. D. G. (1998). Plant disease-resistance proteins and the gene-for-gene concept. *Trends Biochem Sci, 23*(12), 454-456. doi:10.1016/S0968-0004(98)01311-5

van der Hoorn, R. A. L., & Kamoun, S. (2008). From Guard to Decoy: A new model for perception of plant pathogen effectors. *Plant Cell, 20*(8), 2009-2017. doi:10.1105/tpc.108.060194

Vaughn, J. N., & Bennetzen, J. L. (2014). Natural insertions in rice commonly form tandem duplications indicative of patch-mediated double-strand break induction and repair. *Proc Natl Acad Sci U S A, 111*(18), 6684-6689. doi:10.1073/pnas.1321854111

Vrana, J., Kubalakova, M., Cihalikova, J., Valarik, M., & Dolezel, J. (2015). Preparation of sub-genomic fractions enriched for particular chromosomes in polyploid wheat. *Biologia Plantarum, 59*(3), 445-455. doi:10.1007/s10535-015-0522-1

Vrana, J., Kubalakova, M., Simkova, H., Cihalikova, J., Lysak, M. A., & Dolezel, J. (2000). Flow sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L.). *Genetics, 156*(4), 2033-2041.

Wang, J., Luo, M. C., Chen, Z., You, F. M., Wei, Y., Zheng, Y., & Dvorak, J. (2013). *Aegilops tauschii* single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. *New Phytol, 198*(3), 925-937. doi:10.1111/nph.12164

Wang, M. N., Wan, A. M., & Chen, X. M. (2015). Barberry as alternate host is important for *Puccinia graminis* f. sp *tritici* but not for *Puccinia striiformis* f. sp *tritici* in the US Pacific Northwest. *Plant Dis, 99*(11), 1507-1516. doi:10.1094/Pdis-12-14-1279-Re

Wang, Q., Liu, Y., He, J., Zheng, X., Hu, J., Liu, Y., . . . Wan, J. (2014). *STV11* encodes a sulphotransferase and confers durable resistance to rice stripe virus. *Nat Commun, 5*, 4768. doi:10.1038/ncomms5768

Wang, S. C., Wong, D. B., Forrest, K., Allen, A., Chao, S. M., Huang, B. E., . . . Sequencing, I. W. G. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol J, 12*(6), 787-796. doi:10.1111/pbi.12183

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet, 10*(1), 57-63. doi:10.1038/nrg2484

Webb, C. A., & Fellers, J. P. (2006). Cereal rust fungi genomics and the pursuit of virulence and avirulence factors. *FEMS Microbiol Lett, 264*(1), 1-7. doi:10.1111/j.1574-6968.2006.00400

Wicker, T., Buchmann, J. P., & Keller, B. (2010). Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res, 20*(9), 1229-1237. doi:10.1101/gr.107284.110

Wicker, T., Buchmann, J. P., & Keller, B. (2010). Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res, 20*(9), 1229-1237. doi:10.1101/gr.107284.110

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R. H., . . . Choulet, F. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genom Biol, under review*.

Wicker, T., Yu, Y. S., Haberer, G., Mayer, K. F. X., Marri, P. R., Steve, R. W., . . . Roffler, S. (2016). DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. *Nat Commun, 7*,12790.

Williams, S. J., Sohn, K. H., Wan, L., Bernoux, M., Sarris, P. F., Segonzac, C., . . . Jones, J. D. (2014). Structural basis for assembly and function of a heterodimeric plant immune receptor. *Science, 344*(6181), 299-303. doi:10.1126/science.1247357

Woodhouse, M. R., Schnable, J. C., Pedersen, B. S., Lyons, E., Lisch, D., Subramaniam, S., & Freeling, M. (2010). Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol, 8*(6). doi:10.1371/journal.pbio.1000409

Woodhouse, M., Pedersen, B., & Freeling, M. (2010). Transposed genes in *Arabidopsis* are often associated with flanking repeats. *PLoS Genet 6(5):e1000949.*

Woolhouse, M. E. J., Webster, J. P., Domingo, E., Charlesworth, B., & Levin, B. R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat Genet, 32*(4), 569-577. doi:10.1038/ng1202-569

Wulff, B. B., & Moscou, M. J. (2014). Strategies for transferring resistance into wheat: from wide crosses to GM cassettes. *Front Plant Sci, 5*, 692. doi:10.3389/fpls.2014.00692

Yahiaoui, N., Brunner, S., & Keller, B. (2006). Rapid generation of new powdery mildew resistance genes after wheat domestication. *Plant J, 47*(1), 85-98. doi:10.1111/j.1365-313X.2006.02772

Yang, Y., Sterling, J., Storici, F., Resnick, M. A., & Gordenin, D. A. (2008). Hypermutability of damaged single-strand dna formed at double-strand breaks and uncapped telomeres in yeast *Saccharomyces cerevisiae*. *Plos Genet, 4*(11). doi:e100026410.1371/journal.pgen.1000264

Zapata, L., Ding, J., Willing, E. M., Hartwig, B., Bezdan, D., Jiao, W. B., . . . Schneeberger, K. (2016). Chromosome-level assembly of *Arabidopsis thaliana Ler* reveals the extent of translocation and inversion polymorphisms. *Proc Natl Acad Sci U S A, 113*(28), E4052-4060. doi:10.1073/pnas.1607532113

Zimin, A. V., Puiu, D., Hall, R., Kingan, S., Clavijo, B. J., & Salzberg, S. L. (2017). The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience, 6*(11), 1-7. doi:10.1093/gigascience/gix097

Zuo, W., Chao, Q., Zhang, N., Ye, J., Tan, G., Li, B., . . . Xu, M. (2015). A maize wall-associated kinase confers quantitative resistance to head smut. *Nat Genet, 47*(2), 151-157. doi:10.1038/ng.3170

Acknowledgements

The last three and a half years of my PhD have been a great learning experience which has helped me to develop analytical thinking and has improved my scientific skills. I would like to take this oppurtunity to thank everyone who has helped me scientifically or emotionally during my PhD.

Firstly, I would like to express my sincere gratitude to my supervisor **Asst. Prof. Simon Krattinger** for the continuous support of my PhD study, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my PhD study. Thank you for always helping me out with everything. I highly appreciate your support and assistance.

I am grateful to **Prof. Dr. Beat Keller** for giving me this great opportunity to work in his lab and for his insightful comments and encouragement. I appreciate the time you spent for helping me with the thesis writing. Thanks are also due to **Dr. Thomas Wicker** for introducing me to his world of Bioinformatics and for his Perl scripts. You have always been very supportive. *Thank you so much!* I am also thankful to **Prof. Dr. Ueli Grossniklaus** for being my committee member and for his valuable suggestions.

I am also thankful to our collaborators, **Dr. Jaroslav Dolezel** and **Hana Simkova** for codesigning the *Lr22a* gene cloning projects; **Dario Fossati, Odile Moullet** and **Cécile Brabant** for generating the 'CH Campala *Lr22a*'; **Dovetail genomics** for producing wonderful assembly of the Chromosome 2D of 'CH Campala *Lr22a*'; **Manuel Spannagl**, **Marius Felder** and **Thomas Lux** for doing the gene annotation of the 'CH Campala *Lr22a*' 2D chromosome; **Burkhard Steuernagel** and **Brande Wulff** for the NLR annotation.

I would also like to thank **Geri** and **Helen** for their technical assistance and valuable advices. Big thanks to **Christian**, **Kari**, **Linda** and **Esther** for their immense support with

green house and field. I am also thankful to the librarian **Martin Spinnler** for his help in arranging the papers whenever I needed.

I am very thankful to my dear friend and flatmate, **Shibu** for always being there for me. You have been a great support emotionally and I will cherish all our trips and fun we have had in the last three and a half years. I would like to thank my dear friend **Jyoti** for the stimulating discussions and immense support. Thank you for always listening to me. I am also very thankful to my fellow labmates and friends, **Markus** for proving me with the leaf rust isolates; **Stephanie** and **Julia** for afternoon walks and most importantly thank you Stephanie for the german translation of my thesis summary; **Javi** for being a good listener and a great friend, **Stephan**, **Fabrizio** and **Manuel** for their support with Bioinformatics. Big thanks to my master student **Patrick** for his good work and contribution to the *Lr22a* project. Thanks a lot to all the other people whom I haven't mentioned but have made my stay in Swiss enjoyable and worth remembering for lifetime.

Finally, I would like to thank my parents, **Sukhpal Singh** and **Surjeet Kaur** for their trust and confidence in me. My brother **Anantdeep** and my sister-in-law, **Ramandeep** for always standing by my side and cheering me up. Thanks to my nephew, **Harnidh**, for his giggles and charm which always made my day.

I would also like to express my appreciation for the continuous patience, love, support and understanding that my fiancé, **Amandeep Sahota** has given me. Thank you for always cheering me up.

Curriculum Vitae

Family name	THIND
First name	Anupriya Kaur
Date of birth	06 May 1991
Place of Origin	Shahjahanpur, UP, India

Education

2008-2011	B.Sc. Biotechnology (Hons.), Panjab University, Chandigarh, India
2011-2013	Masters in Biotechnology, Punjab Agricultural University, Punjab, India Master's supervision: Dr. Parveen Chunneja Master's thesis title: **Sequencing of gliadin genes in old and modern wheat varieties for identification of celiac disease causing epitopes**
2015-2018	PhD in Plant Science, Department of Plant and Molecular Biology, University of Zurich, Switzerland PhD Supervision: Asst. Prof. Dr. Simon Krattinger Title: **Cultivar-specific long-range chromosome assembly permits rapid gene isolation and high-quality comparative analysis in hexaploid wheat**

Publications

Thind AK, Wicker T, Müller T, Ackermann PM, Steuernagel B, Wulff BBH, Spannagl M, Twardziok SO, Felder M, Lux T, Mayer KFX, International Wheat Genome Sequencing Consortium, Keller B, Krattinger SG (2018) Chromosome-scale comparative sequence analysis unravels molecular mechanisms of genome differences between two wheat cultivars. *Genome Biology*. 19:104.

Thind AK, Wicker T, Šimkov. H, Fossati D, Moullet O, Brabant C, Vr.na J, Doležel J and Krattinger SG (2017) Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. *Nature Biotechnology* 35: 793-79.

Thind AK, Wicker T and Krattinger SG (2017) Rapid identification of rust resistance genes through cultivar-specific de novo chromosome assemblies. In: Periyannan S (ed) *Methods in Molecular Biology, springer,* 1659:245-255.

Kaur A, Bains NS, Sood A, Yadav B, Sharma P, Kaur S, Garg M, Midha V and Chhuneja P (2016) Molecular characterization of α-gliadin gene sequences in Indian wheat cultivars vis-.-vis celiac disease eliciting epitopes. *Journal of Plant Biochemistry and Biotechnolnology*. DOI 10.1007/s13562-016-0367-5.